A CRITERION FOR SELECTING THE PROBABILITY DENSITY FUNCTION OF BEST FIT FOR HYDROLOGIC DATA

A THESIS

Presented to

The Faculty of the Division of Graduate Studies

By

Donthamsetti Veerabhadra Rao

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy in the School of Civil Engineering

Georgia Institute of Technology

March, 1978

A CRITERION FOR SELECTING THE PROBABILITY

DENSITY FUNCTION OF BEST FIT FOR HYDROLOGIC DATA

Approved:

James R. Wallace, Chairman

L. Douglas James

Willard M. Snyder

May 1, 1978

Date Approved by Chairman:

ACKNOWLEDGEMENTS

The author wishes to extend his gratitude to all who helped to make this investigation possible.

The author is grateful to his advisor, Dr. J. R. Wallace, whose encouragement, advice and counsel are in a great measure responsible for the completion of this work, and to the members of the Reading Committee, Dr. L. Douglas James and Professor Willard M. Snyder for their invaluable suggestions and comments.

Financial support for this investigation was received from the Agricultural Research Service, Southeast Watershed Laboratory, and from the School of Civil Engineering, Georgia Institute of Technology.

The author wishes to thank his wife Sreedevi and sons, Ravi, Ajay, and Aravind for their patience and cooperation during the period of this work.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF TABLES (continued)

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS (continued)

LIST OF ILLUSTRATIONS (continued)

SUMMARY

The objective of this study was to investigate potential criteria for selecting the probability density function (PDF) of best fit for hydrologic data. The major steps were the examination of various statistical methods for parameter estimation, the determination of differences between PDF's (particularly how the characteristic shapes of the PDF's depend on the value of population variance), the examination of the discrepancies between the sample moments and the moments of the fitted PDF, and the identification of strengths and weaknesses of several potential "best fit" criteria.

The three PDF's examined in this investigation (lognormal, gamma, and Gumbel) were found to have similar shapes for certain values of the population variance but to differ greatly at other values. When fitting a sample to one of these PDF's each of the methods used for parameter estimation ignores certain characteristics of the sample while it emphasizes others. The method of moments (MO) ignores the overall form of the sample distribution (histogram) while fitting the first two moments of the distribution. On the other hand, the method of maximum likelihood (ML) and the method of least squares (LS), each in their own way, fit only the form or shape of the sample and ignore the sample moments. It was concluded that the sample moments will generally be approximately equal to the moments of the fitted PDF when the correct, or parent, PDF is fitted by a shape fitting method such as LS or ML. Thus a criterion for selecting an appropriate probability density function f(x) for fitting a

random sample (X_i) of size n is to select that f(x) which makes the statistic

$$\frac{\text{Var } (X_i | f(x;\theta))}{\text{Var } (X_i)}$$

closest to unity, where

$$\text{Var } (X_i) = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1)$$

$$\text{Var } (X_i | f(x;\theta)) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x;\theta)dx$$

and the parameters of f(x), θ, are estimated by a shape fitting method such as maximum likelihood or least squares. (This criterion is herein referred to as the variance ratio.) Other criteria identified as possible discriminators of PDF's were the statistics of chi-square and Kolomogorov-Smirnov (K-S) goodness-to-fit tests, sum of squared errors of least squares fit.

The variance ratio test which was found to be the most satisfactory criterion for use in identifying the parent PDF, was applied to 67 real hydrologic samples (annual peak flows) and a 'best fit' to either a LN, GA or GU PDF was determined. The GU was never found to be unequivocally superior to both GA and LN. Nineteen samples were judged by a criterion outlined in the study, to contain outliers. Neither the LN or GA distributions provided acceptable fits for four samples.

Throughout the study it was noted that errors in predictions introduced in frequency analysis by not choosing the 'best' probability density function are larger when computations are made by the maximum likelihood or least squares method than when the computations are made by the method of moments.

CHAPTER I

INTRODUCTION

Measurements of hydrological events, such as precipitation and streamflow, provide the raw data for quantitative hydrologic analysis. One analytic approach is to use the data for quantitative expression of hydrologic processes in deterministic models. An alternative is to treat the measured hydrologic phenomena as stochastic processes, i. e., processes governed by the laws of chance. With this approach, methods based on the theories of probability and statistics can be used to analyze these hydrologic data systematically and draw inferences from the data.

One of the chief applications of this approach is to estimate the probability of occurrence of hydrologic events, the frequencies of phenomena such as floods, droughts, storages, rainfalls, water qualities and waves. Although extensive work, both theoretical and applied, has been done to develop a methodology for quantifying the frequency of hydrologic events (see 137 references given by Chow, 1964), many issues still remain elusive and intriguing. At a recent International Symposium in Hydrology (Schulz, E. F., et al., 1973), aspects of flood frequency analysis were discussed at length, and the session chairman's concluding remarks were ". . . [the proceedings show] that there is no agreed or universally applicable approach for analysis of flood probabilities." Since no universally accepted method has been demonstrated, the method of analysis continues to be based on engineering judgment and intuition.

where $f(x_i)$ is the PDF evaluated at the sample values, x_i, and n is the number of items in the sample (a detailed description of ML method is presented in Chapter II). Since the values of $f(x_i)$ represent the sahpe of the PDF, it can be stated that the parameters are selected by determining the shape of the PDF that best fits the sample data, where "best fit" is determined by maximizing the probability of the sample. Thus, the ML method does not insure that the moments of the fitted PDF are equal to the sample moments. In the LS method, the parameter estimates are obtained by minimizing the sum of squared errors (SSE) between observed frequencies of a histogram formed from a data sample and the frequencies based on the selected PDF. This procedure is essentially equal to fitting the shape of a selected PDF to the shape of a histogram of the sample, and will not insure that the sample moments and the moments of the fitted PDF are equal. (An extension of the LS method is the method of minimum chi-square (MCS) in which the objective is to minimize a sum of weighted errors where the weight is a function of the PDF.) The above analysis shows that each of the three statistical estimation methods, MO, ML and LS, in fact, fits only some of the characteristics of a sample; the MO method fits the sample by matching only the moments of the selected PDF and the ML and LS methods fit the sample by matching only the shape of the selected PDF. Consideration of this basic distinction suggests that it may be possible to find a PDF which when fit by the MO method matches shape as well as moments or when fit by the ML or LS method matches moments as well as shape. Accordingly, one of the major objectives of this research became to test the hypothesis that an

appropriate PDF to be used for fitting a data sample can be selected by choosing that PDF which approximates the shape of the sample distribution and, at the same time, has its moments approximately equal to the moments of the sample. The approach to test this hypothesis, as well as to examine certain other criteria for the selection of appropriate PDF's for fitting data samples, was through a systematic study of three PDF's in general use in hydrologic analysis, namely, the lognormal (LN), the gamma (GA), and the Gumbel (GU), each a two-parameter PDF.

The method of moments is quite simple. The ML equations are too complex to solve in some cases, and LS[1] method has, in the past, been difficult to apply to nonlinear PDF's. It was only recently that a method of applying LS technique to estimate parameters of PDF's was developed by Snyder (1972).

Outline of Study

The plan of study was first to attempt to understand the subtle differences between the LN, GA and GU PDF's in their moments and in their characteristic shapes (a dimensionless approach was adopted for this purpose). From this knowledge, specific discriminating criteria were developed for choosing a PDF to fit a sample. To test the validity of these discriminating criteria and to quantify certain intuitive expectations as to what would happen if data from one PDF

---

[1]The original intent of this research was a detailed investigation of frequency analysis by the method of least squares. After considerable study on this topic was completed, the emphasis changed to that of examining goodness-of-fit criteria. However, the results of the study on LS fitting were preserved and are presented in Appendix A.

were fit by assuming another PDF, a systematic study based on numerical simulation experiments was made.

The presentation follows the above outline. Chapter II presents relevant theoretical aspects of statistical parametric estimation methods. Chapter III discusses the characteristic shapes of three families of PDF's (LN, GA, and GU). Chapter IV describes some criteria to discriminate PDF's and outlines numerical experimentation undertaken to test the validity of selected PDF discriminating criteria. Chapter V gives detailed accounts of the various numerical experiments and their results. The numerical experiments consisted of generating synthetic samples and fitting the samples to different PDF's by MO, ML, LS, and MCS. Chapter VI recounts the results of fitting a large number of long recorded sequences of annual floods in all parts of the United States to selected PDF's. Chapter VI also presents an analysis of the influence of outliers on fits made by various estimation methods and an account of how frequency analysis is affected by larger samples (longer period of record). Chapter VII summarizes the results of this study and lists certain conclusions.

CHAPTER II

ESTIMATION OF PARAMETERS OF PROBABILITY DISTRIBUTIONS

Estimation is the process of extracting quantitative information on a parameter or signal function from noise-corrupted observations (Nahi, 1969). While estimation techniques are extensively dealt with in standard works on inferential statistics (Kendall and Stuart (1973), Cramer (1945), Von Mises (1964), Graybill (1961), Rao (1965) among others) and monographs (Nahi (1969), Wassan (1970), Bard (1974)), the important theoretical concepts for the present study are collected and presented in this Chapter.

The properties of estimators are usually described by the following terminology: (Hines and Montgomery, 1972)

Unbiasedness: An estimator $\hat{\theta}$ is said to be an unbiased estimator of θ if $E(\hat{\theta}) = \theta$, where E denotes mathematical expectation.

Consistency: An estimator $\hat{\theta}_n$ (more accurately, a sequence of estimators $\{\hat{\theta}_n\}$) is said to be consistent for θ if the limit of the probability that $|\hat{\theta}_n - \theta| < \varepsilon$ is 1 as n approaches infinity. $\hat{\theta}_n$ is the estimate of θ based on a sample of size n.

Efficiency: If $\hat{\theta}$ and θ^* are unbiased estimators of θ, then $\hat{\theta}$ is relatively more efficient than θ^* if $V(\hat{\theta}) < V(\theta^*)$, where V denotes variance.

One general procedure of frequency analysis essentially consists of representing the data, in its original form or by a suitable transformation, by a probability density function (PDF). Markovic (1965) fitted five PDF's, namely, normal, lognormal (with two and three parameters), and gamma (with two and three parameters) to distributions of annual precipitation and annual runoff in the western United States and southwestern Canada. Based on a chi-square goodness-of-fit test, he concluded that all five PDF's studied were applicable and none was more suitable than any other.

This inability to choose on the basis of statistical validity, however, is not very useful to the engineering hydrologist because of the differing results obtained when the same data sample is fitted to different PDF's. Flood frequency analyses conducted by Cruff and Rantz (1965) indicated that four statistical distributions, lognormal, Gumbel, gamma (all two-parameter) and Pearson type III, when applied to the same peak discharge data, gave widely differing results for individual stations. All these distributions have been extensively used and each has the support of reputable statistician-hydrologists as being the distribution that best describes the occurrence of hydrologic events (Cruff and Rantz, 1965). One may infer from this situation either that the distribution that truly describes the occurrence of hydrologic events is not known or that no single distribution covers all of the many widely varying hydrological conditions found at different places. Nevertheless, for interagency consistency in water resources planning, the U. S. Water Resources Council has recommended universal use by Federal agencies of the

log Pearson type III distribution (Bulletin No. 15, see Bibliography).

While many density functions are available and applicable, one can (with the specifics depending on the nature of the hydrologic variables being examined) eliminate many density functions as unsuitable for hydrologic frequency analysis leaving only a limited number which may be considered suitable. The distributions which are extensively applied to hydrologic data are normal, lognormal, gamma (particularly its form with two parameters and Pearson type III), double exponential and simple exponential functions of extreme values (Yevjevich, 1972).

The characteristic shape of a PDF is determined by the general algebraic form of the function and by the values of its parameters, and the value for a parameter estimated from a sample will depend on the method used for the estimation. The three principal methods are the method of moments (MO), the method of maximum likelihood (ML), and the method of least squares (LS). In the MO method, m population moments of the selected PDF are set equal to m corresponding sample moments, where m is the number of parameters in the PDF. This procedure provides m equations with m unknown parameters. These equations, when solved, yield the required parameter estimates. In the ML method, parameter estimates for a selected PDF are obtained by maximizing a function (Hines and Montgomery, 1972)

$$L(\Theta) = f(x_1)f(x_2)\ldots f(x_n)$$

Minimum Variance: An estimator $\hat{\theta}$ is said to be a minimum variance estimator of θ if $E[\hat{\theta} - E(\hat{\theta})]^2 \leq E[\theta^* - E(\theta^*)]$, where θ^* is any other estimator for θ.

Invariance: An estimator $\hat{\theta}$ of θ is said to be an invariant estimator for a certain class of transformations g if the estimator is $g(\hat{\theta})$ when the transformation changes the parameter to $g(\theta)$.

The Sample, its Moments and its Distribution

Let $f(x)$ and $f_0(x)$ be two PDF's. Let $x_1, x_2, \ldots, x_n$ be a random sample from $f_0(x)$. Let $f_0(x)$ and $f(x)$ be evaluated at the x_i's from the sample. If, by some criterion of closeness, it can be determined that the $f(x_i)$'s are close to the values $f_0(x_i)$, then it may be stated that the shapes of $f_0(x)$ and $f(x)$ are similar. For example, curves (a) and (b) in Figure 2.1 might be considered similar while (c) and (d) might be dissimilar.

It is also of interest to compare the moments of $f_0(x)$ to the moments of $f(x)$. The first moment about the origin, the mean ($\bar{x}$), and the second moment about the mean, the variance (S^2) of the sample, are given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad (2.1)$$

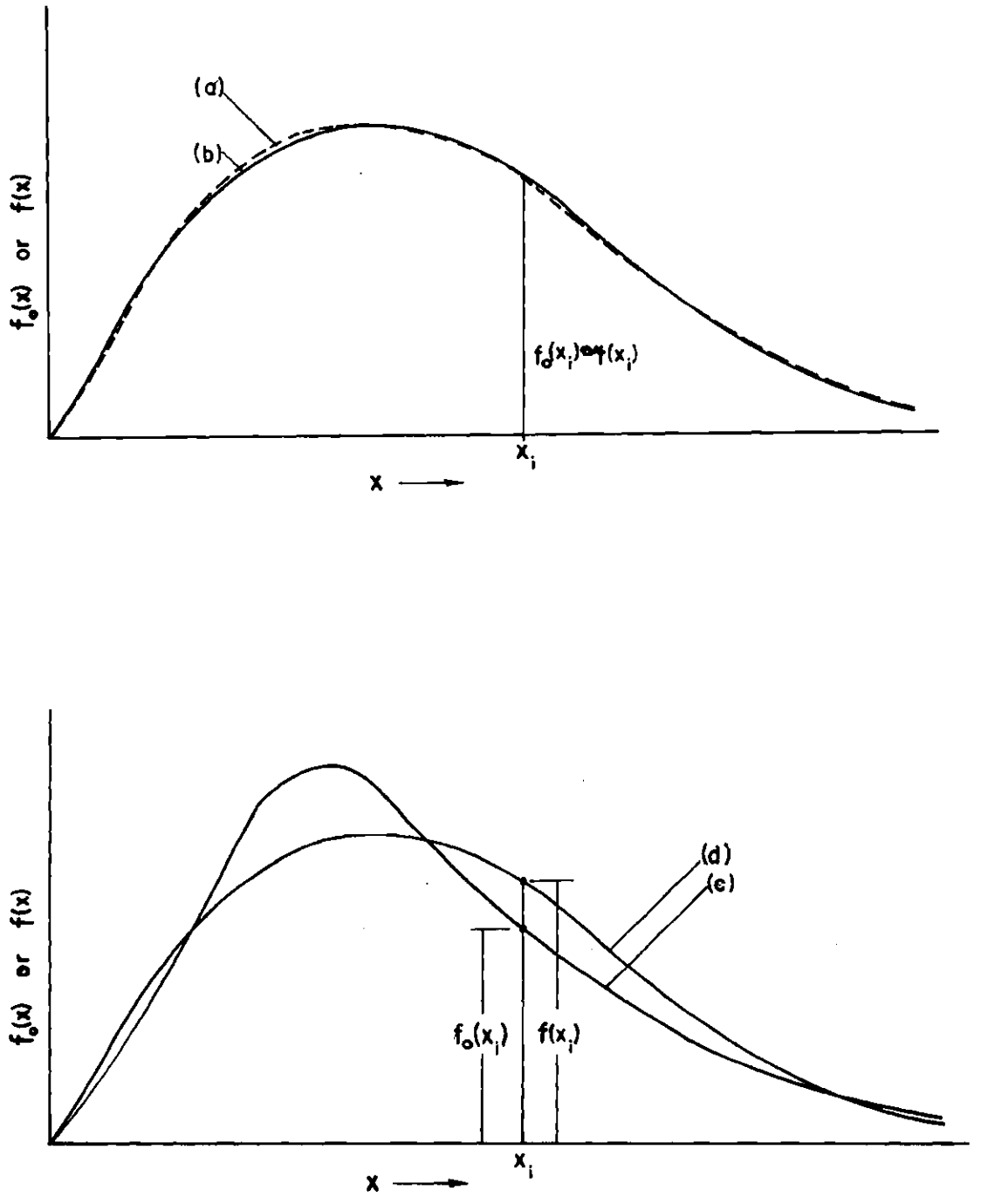


Figure 2.1 Probability Density Curves of $f_o(x)$ and $f(x)$

and

$$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \tag{2.2}$$

Other higher sample moments may be similarly formulated. Also, the first moment about the origin, the mean (μ) and the second moment about mean, the variance (σ^2), of f(x) are given by

$$\mu = \int_{-\infty}^{\infty} xf(x)dx \tag{2.3}$$

$$\sigma^2 = \int_{-\infty}^{\infty}(x - \mu)^2 f(x)dx \tag{2.4}$$

Other higher moments of f(x) may be similarly formulated.

If, by some criterion of closeness, it can be shown that

$$\bar{x} \simeq \int xf(x)dx \tag{2.5}$$

and

$$S^2 \simeq \int (x - \mu)^2 f(x)dx \tag{2.6}$$

then it may be said that the moments of $f_0(x)$ and f(x) are similar.

Parametric Estimation Methods

The parametric estimation methods used in this study are: (1) the method of moments (MO), (2) the method of maximum likelihood (ML),

(3) the method of least squares (LS), and (4) the method of minimum chi-square (MCS). Brief descriptions of these methods are given below.

The Method of Moments (MO)

The MO method was first proposed by Karl Pearson in 1894 (Hines and Montgomery, 1972) and offers a very simple procedure in most cases. This method consists essentially of equating m sample moments to m corresponding population moments, where m is the number of parameters in PDF.

Let $f(x;\theta)$ be a density function characterized by an unknown parameter θ. The first population moment about zero is

$$\mu = \int_{-\infty}^{\infty} xf(x;\theta)dx \tag{2.7}$$

which will, in general, be a function of the unknown parameter θ. Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from the density $f(x)$, and define the first sample moment to be

$$\bar{x} = m_1' = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.8}$$

Upon equating 2.7 and 2.8, one obtains

$$\mu = m_1'$$

from which θ can be estimated. The procedure will easily generalize to m unknown parameters, $\underset{\sim}{\theta}$. (Note that $\underset{\sim}{\theta}$ are not the statistical

parameters, they are the parameters in the PDF.) In this case, one needs the first m population moments about zero

$$\mu_t' = \int_{-\infty}^{\infty} x^t f(x;\underset{\sim}{\theta})dx, \qquad t = 1, 2, \ldots, m \tag{2.9}$$

and the first m sample moments

$$m_t' = \frac{1}{n}\sum_{i=1}^{n} x_i^t, \qquad t = 1, 2, \ldots, m \tag{2.10}$$

Equating 2.9 and 2.10 yeilds the m simultaneous equations involving the m unknown parameters

$$\mu_t' = m_t', \qquad t = 1, 2, \ldots, m \tag{2.11}$$

The solution to Equation 2.11 yields estimates of the $\underset{\sim}{\theta}$ by the method of moments.

In the foregoing procedure, if the sample is from an unknown density and f(x) is a hypothesized density, the method of moments merely uses the knowledge of the sample moments and assigns to the sample the hypothesized distribution with population moments equal to the sample moments. Thus, the moment estimates are not necessarily optimal in any sense.

The Method of Maximum Likelihood (ML)

The ML method of generating estimators of unknown parameters was first introduced by Fisher (1921). This method generally leads to efficient and consistent estimators. ML estimators are not always

unbiased, but in many cases they may be easily modified to make them unbiased. Also, the ML estimators are invariant. In general, ML estimators are known to have very desirable statistical properties (Kendall and Stuart, 1973).

Let x be a random variable having probability density function f characterized by m unknown parameters $\underset{\sim}{\theta}$. Let $x_1, x_2, \ldots, x_n$ be the observed values in a ramdom sample of n. Then the likelihood function of the sample is defined to be

$$L(x_1,x_2, \ldots,x_n;\underset{\sim}{\theta}) = f(x_1;\underset{\sim}{\theta})f(x_2;\underset{\sim}{\theta}) \ldots f(x_n;\underset{\sim}{\theta}) \qquad (2.12)$$

The ML estimator of $\underset{\sim}{\theta}$ is defined as the value $\underset{\sim}{\hat{\theta}}$ such that

$$L(x_1,x_2, \ldots,x_n;\underset{\sim}{\hat{\theta}}) \geq L(x_1,x_2, \ldots,x_n;\underset{\sim}{\theta}) \qquad (2.13)$$

That is, the maximum likelihood estimate of $\underset{\sim}{\theta}$ is the value, say $\underset{\sim}{\hat{\theta}}$, that maximizes the likelihood function when it is considered as a function of $\underset{\sim}{\theta}$. In other words, the ML estimate maximizes the probability of the joint occurrence of the sample results, $f(x_i)$. This is an example of fitting the shape of the sample distribution (see Figure 2.1) and is thus intuitively an appealing approach. It may, however, be noted that the ML method maximizes the probability of the joint occurrence of the $f(x_i)$ and, thus, will fit only the 'overall' shape of the sample distribution.

The ML estimating procedure consists essentially of formulating the right hand side of Equation 2.12 for a given PDF and equating to

zero its derivatives with respect to each parameter.

If the sample mentioned in the foregoing discussion is from an unknown distribution and the frequency function $f(x;\underset{\sim}{\theta})$ is a hypothesized distribution, the ML estimation method would present a 'fit' from the family of the 'hypothesized' density that would approximate the sample distribution. However, the ML method does not insure that the moments of the fitted PDF match the moments of the sample. The ML estimators, nevertheless, accomplish an important property desired in the optimal solution, namely, fitting the shape of the sample distribution. The moments of the fitted PDF may be easily computed from the knowledge of ML estimators of $\underset{\sim}{\theta}$ and the moments of the sample may be compared with the moments of the fit.

The Method of Least Squares (LS)

The method of least squares (due to Gauss) is one of the oldest procedures used in inferential statistics. It produces estimators that, in many important cases, are unbiased and consistent and in some cases minimum variance unbiased (Hines and Montgomery, 1972).

Let an observation consist of two n-tuples of known data $\underset{\sim}{x}_n = (x_1, x_2, \ldots, x_n)$ and $\underset{\sim}{y}_n = (y_1, y_2, \ldots, y_n)$ and let the n-tuple $\underset{\sim}{y}_n$ be related to n-tuple $\underset{\sim}{x}_n$ by equation

$$Y_i = f(x_i;\underset{\sim}{\theta}) = E(y|x_i) \tag{2.14}$$

where the form of f is known and $\underset{\sim}{\theta}$ represents a vector of m parameters.

In the method of least squares the elements which compose the random observation (of size n) are assumed to be of the form

$$y_i = f(x_i;\theta) + e_i \qquad i = 1, 2, \ldots, n \tag{2.15a}$$

or

$$y_i = Y_i + e_i \tag{2.15b}$$

in which the parameters $\underset{\sim}{\theta}$ are unknown. It is desired to estimate $\underset{\sim}{\theta}$. The values e_i, called residuals, may be thought of as "random errors" which obscure the true values of the y_i's. The least squares estimates of $\underset{\sim}{\theta}$ are those values of $\underset{\sim}{\theta}$ which minimize the sum of squared errors

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - Y_i)^2 \tag{2.16}$$

In the most general form a weight w may be assigned to each squared error [Levenberg (1944), Grant (1973)] and the weighted residuals and the weighted sum of squared errors is given as:

$$e_i = w_i^{1/2} (y_i - Y_i) \tag{2.17}$$

$$SSE(\underset{\sim}{\theta}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} w_i (y_i - Y_i)^2 \tag{2.18}$$

The weight w may be a function of x, $\underset{\sim}{\theta}$, or both.

The least squares estimators, say $\hat{\underset{\sim}{\theta}}$, will be the solution to the m simultaneous equations

$$\frac{\partial(SSE)}{\partial \underset{\sim}{\theta}} = 0 \qquad (2.19)$$

To use the method of least squares it is not necessary that the probability distribution of either the observations or the random errors be known. Nevertheless, a knowledge of the distribution of the residuals leads to better estimation of the properties of least squares estimators (see Kendall and Stuart, 1973, or Graybill, 1961).

When the model given by Equation 2.15 is linear or intrinsically linear (an intrinsically linear model is one in which the model is made linear by a suitable transformation, frequently by a logarithmic transformation) in parameters $\underset{\sim}{\theta}$ the estimates given by Equations 2.19, known as normal equations, are relatively easy to solve (Graybill (1961) Draper and Smith (1966) among others). If it is impossible to convert a model into a form linear in the parameters the model is said to be intrinsically non-linear (Draper and Smith, 1966). The three probabilistic models used in this study, namely, lognormal, gamma, and Gumbel, are examples of intrinsically non-linear models. In such models the normal equations will be non-linear and obtaining a solution can be extremely difficult. Iterative methods must be employed in nearly all cases. Appendix A presents in detail a method of estimating parameters of non-linear models by least squares.

Intuitively, the LS method, like the method of ML, may be assumed to fit the shape of the model to the shape of the sample as shown in Appendix A. Hence, the LS method will be said to be a "shape fitting" method.

The Method of Minimum Chi-Square (MCS)

If the weight w_i in Equation 2.18 is given a value equal to $1/Y_i$ the weighted SSE has a chi-square distribution and is given by

$$\chi^2(\underset{\sim}{\theta}) = \sum_{i=1}^{n} (y_i - Y_i)^2/Y_i \qquad (2.20)$$

By minimizing $\chi^2(\underset{\sim}{\theta})$ with respect to $\underset{\sim}{\theta}$ an estimate $\hat{\underset{\sim}{\theta}}$, known as a MCS estimate, is obtained. If used to fit models of probability distributions, (see Appendix A) the method is applicable to continuous distributions with grouped data expressed as average frequencies over class intervals, or to absolute frequencies of discrete distributions. The estimators of both MCS and ML methods are known to have the same asympototic properties (Kendall and Stuart, 1973). When dealing with samples from a continuous distribution, observations must be grouped in order to make use of the MCS method. This constraint is not imposed when the ML method is employed. Therefore, it has become much more common to employ ML methods as compared to MCS when fitting sampled data to probability distributions (Kendall and Stuart, 1973).

The MCS method may be viewed essentially as a method of least squares and, like ML and LS methods, can be considered a shape fitting method. In Appendix A the least squares method is studied in a general fashion by using weight, $w_i = (Y)^{-\phi}$ (see Equation 2.18) in which ϕ is a weight exponent. When the value of ϕ is unity the least squares method, as studied in Appendix A, is equivalent to the method of minimum chi-square.

Illustrative Examples of Fitting by Different Methods

In order to illustrate the fitting methods outlined above and the variation in results, this section presents the results obtained in fitting two data samples by MO, ML, LS and MCS to a two parameter lognormal (LN) and to a two parameter gamma (GA) distribution. Table 3.1 gives the form of LN and GA PDF's. Appendix A gives a detailed description of LS method (MCS is a special case of weighted LS). Appendices B and C give MO and ML equations for parameter estimates and the equations needed for LS solution for the LN and GU PDF's, respectively.

The data, which are designated as K_i, are annual series.

Sample 1: Sample size = 100 $y = \ln K$

Sample Mean, $\overline{K} = 0.985$

Sample Unbiased Estimate of Variance, $S_k^2 = 0.628$
(Biased Estimate = 0.622)

Data:

.28	.09	5.51	.66	.88	.51	.86
.23	1.42	2.74	1.08	1.25	.34	.77
2.44	.15	.46	.60	.39	1.44	1.63
.68	1.19	.20	2.52	.58	.94	.30
1.32	.86	.39	.97	1.49	1.49	1.22
.52	.81	.75	.43	.83	2.00	1.18
.57	2.33	.51	.38	.70	.65	1.76
.57	.56	.97	.27	.24	.73	.61
.54	1.54	1.76	.14	.63	.59	.50
.79	.47	.35	1.15	.93	2.44	.47
.18	.71	1.15	1.78	.63	1.10	
.66	1.36	1.22	.52	.73	.72	
.66	.77	.98	.97	.85	.82	
.57	1.32	1.54	.77	.52	1.45	
.33	1.41	1.77	.55	4.12	1.27	

a) Results of Lognormal Analysis

	MO	ML	LS	MCS
Parameter estimate of μ_y, $\hat{\mu}_y$	-.263	-.261	-.229	-.248
Parameter estimate of σ_y, $\hat{\sigma}_y$	.704	.712	.599	.711
Mean of the fitted PDF, μ_F	.985	.992	.952	1.005
Variance of the fitted PDF, σ_F^2	.622	.648	.391	.664
Estimate of 100-year event, K_{S100}	3.954	4.032	3.202	4.076

(The mean and variance of fitted PDF are computed using estimated parameters. See columns (3) and (4) of Table 3.1 for the equations of mean and variance of LN, GA PDF's).

The above table shows that the mean and variance of the fitted PDF, μ_F and σ_F^2, are equal to the sample values for MO (note that in the MO method the biased estimate of sample variance equals σ_F^2) and approximately equal to the sample values for ML and MCS. This result may be expected in MO, but this result in ML and MCS indicates that the ML and MCS methods, while fitting the shape of the sample have also approximately fitted the moments of the sample. The above table also shows that the parameter estimates ($\hat{\mu}_y$ and $\hat{\sigma}_y$) by MO are approximately equal to the parameter estimates of ML and MCS. This result shows that the MO method, while fitting the moments of the sample, also fitted the shape of the sample.

The results of LS in the above table differed from the results of MO, ML and MCS. The LS method used in this study was found to produce biased results sometimes (see Appendix A). This bias could be

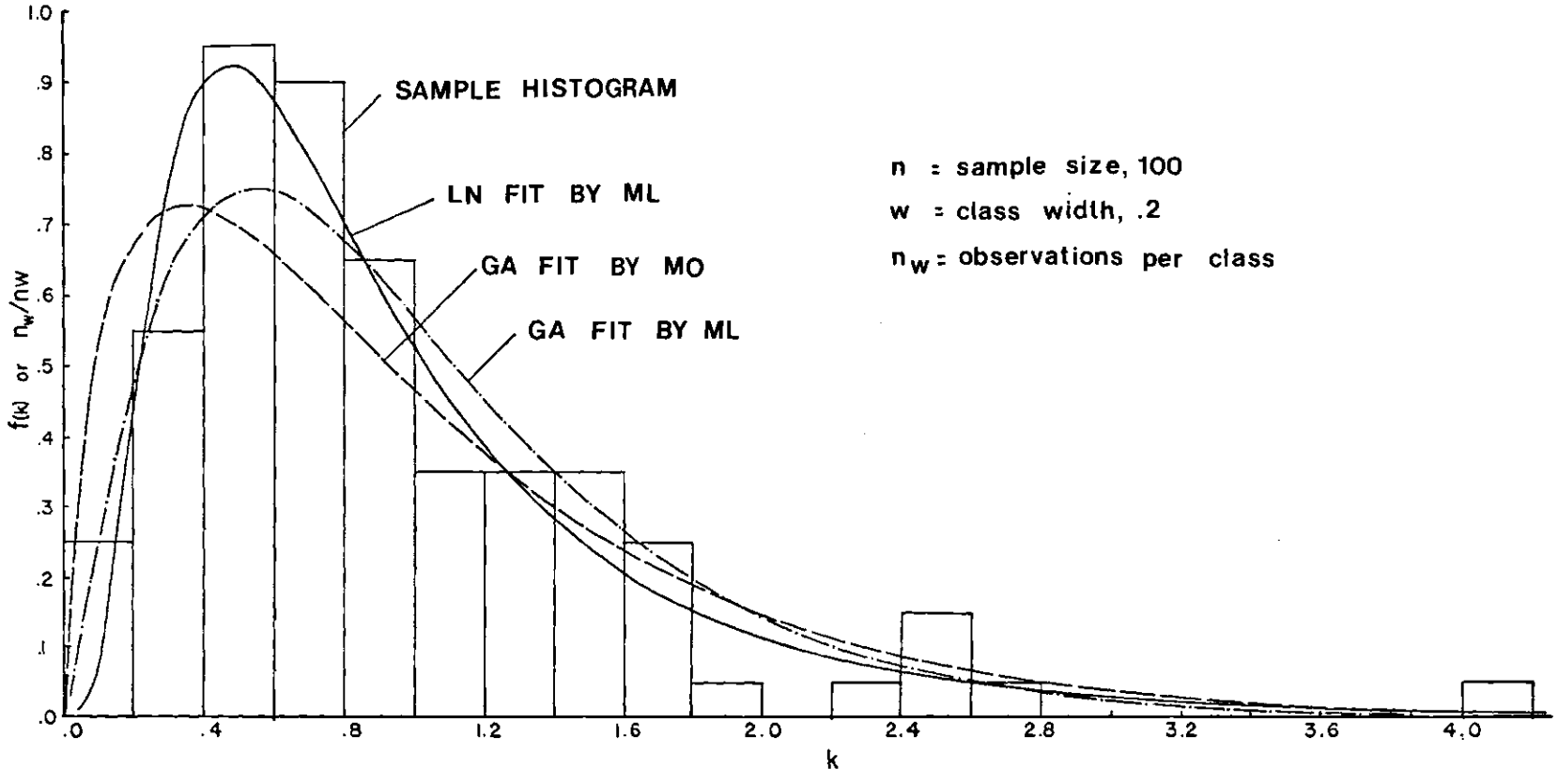


Figure 2.2 Sample 1 from Illustrative Example

corrected by assigning a weight to error terms in LS method. MCS method is a special case of weighted LS method. The LS results in the above table may be viewed essentially as biased results due to some procedural deficiency and thus disregarded. A complete discussion as to how to apply the generalized LS method in fitting PDF's is included in Appendix A.

Note that the values of K_{S100} are approximately equal in MO, ML and MCS methods.

b) Results of Gamma Analysis

	MO	ML	LS	MCS
Parameter estimate of C, $\hat{C}$	1.584	2.216	3.449	2.129
Parameter estimate of D, $\hat{D}$	1.560	2.183	3.006	2.103
Mean of the fitted PDF, μ_F	0.985	1.015	0.873	1.009
Variance of the fitted PDF, σ_F^2	0.622	0.465	0.253	0.479
Estimate of 100-year event, K_{S100}	3.730	3.148	2.440	3.212

The above table shows that the mean and variance of the fitted PDF are equal to the sample values for MO, but differ from the sample values for ML, LS and MCS. In particular, σ_F^2 differed greatly from S_k^2 for ML, LS and MCS. This shows that the ML, LS and MCS did not fit the moments of the sample. The above table also shows that the parameter estimates of MO differed greatly from the parameter estimates of ML, LS and MCS. This shows that the MO method did not fit the shape of the sample.

The above results show that when GA is applied to Sample 1, each of the four statistical estimation methods, MO, ML, LS and MCS, fitted either the moments or the shape, but not both. Note that K_{S100} by MO

differs greatly from K_{S100} by ML, LS and MCS.

For this sample Figure 2.2 shows the sample histogram, LN fit by ML, GA fit by MO and GA fit by ML.

Sample 2: Sample Size = 100

Sample Mean, $\bar{K} = 0.958$

Sample Variance (Unbiased Estimate), $S_k^2 = 0.334$
(Biased Estimate = .331)

Data:

1.01	1.54	.94	.38	1.25	.72	.47
.58	.60	.58	.59	.95	2.13	2.37
.91	1.12	.64	.43	.94	.62	1.25
.22	.61	.56	3.23	.53	.54	1.17
.51	.88	.98	.30	1.45	.94	.81
1.12	.97	.97	1.31	.24	.67	.93
.45	2.23	.55	2.19	1.82	1.20	.47
.56	.87	2.36	.76	1.67	.93	1.55
.48	.67	.89	.73	1.17	.82	.36
.79	.25	1.28	.35	.93	1.31	1.77
.65	.85	.55	1.94	.13	.76	
1.83	1.21	.59	.60	2.04	1.19	
.37	.49	.39	.21	1.28	.92	
.51	.65	1.76	.66	1.67	.16	
1.16	1.62	1.15	.84	1.14	.14	

a) Results of Lognormal Analysis

	MO	ML	LS	MCS
$\hat{\mu}_y$	-.197	-.227	-.138	-.178
$\hat{\sigma}_y$	.555	.638	.638	.647
μ_F	.958	.977	1.068	1.032
σ_F^2	.331	.480	.573	.553
K_{S100}	2.986	3.517	3.842	3.775

The above table shows that μ_F and σ_F^2 are equal to the sample values for MO, but they differ from the sample values for ML, LS and MCS. Also, the parameter estimates of MO, particularly that of $\hat{\sigma}_y$, differed greatly from the parameter estimates of ML, LS and MCS. These results show that the ML, LS and MCS did not fit the moments of the sample and the MO did not fit the shape of the sample. Thus, when LN was applied to Sample 2, the MO, ML, LS and MCS methods fitted only some of the characteristics of the sample, but not all. Note that K_{S100} by MO differs greatly from K_{S100} by ML, LS and MCS.

b) Results of Gamma Analysis

	MO	ML	LS	MCS
$\hat{C}$	2.894	2.987	3.087	2.892
$\hat{D}$	2.773	2.863	2.937	2.841
μ_F	0.958	0.958	0.951	0.982
σ_k^2	0.331	0.321	0.308	0.340
K_{S100}	2.770	2.735	2.688	2.812

The above table shows that σ_F and σ_F^2 are equal to the sample values for MO and they are approximately equal to the sample values. Also, the parameter estimates of MO are approximately equal to the parameter estimates of ML, LS and MCS. These results show that the four statistical methods fitted, in general, all the characteristics of the sample. Note that the values of K_{S100} are approximately equal in all four methods.

For Sample 2, Figure 2.3 shows the sample histogram, GA fit by ML, LN fit by MO and the LN fit by ML.

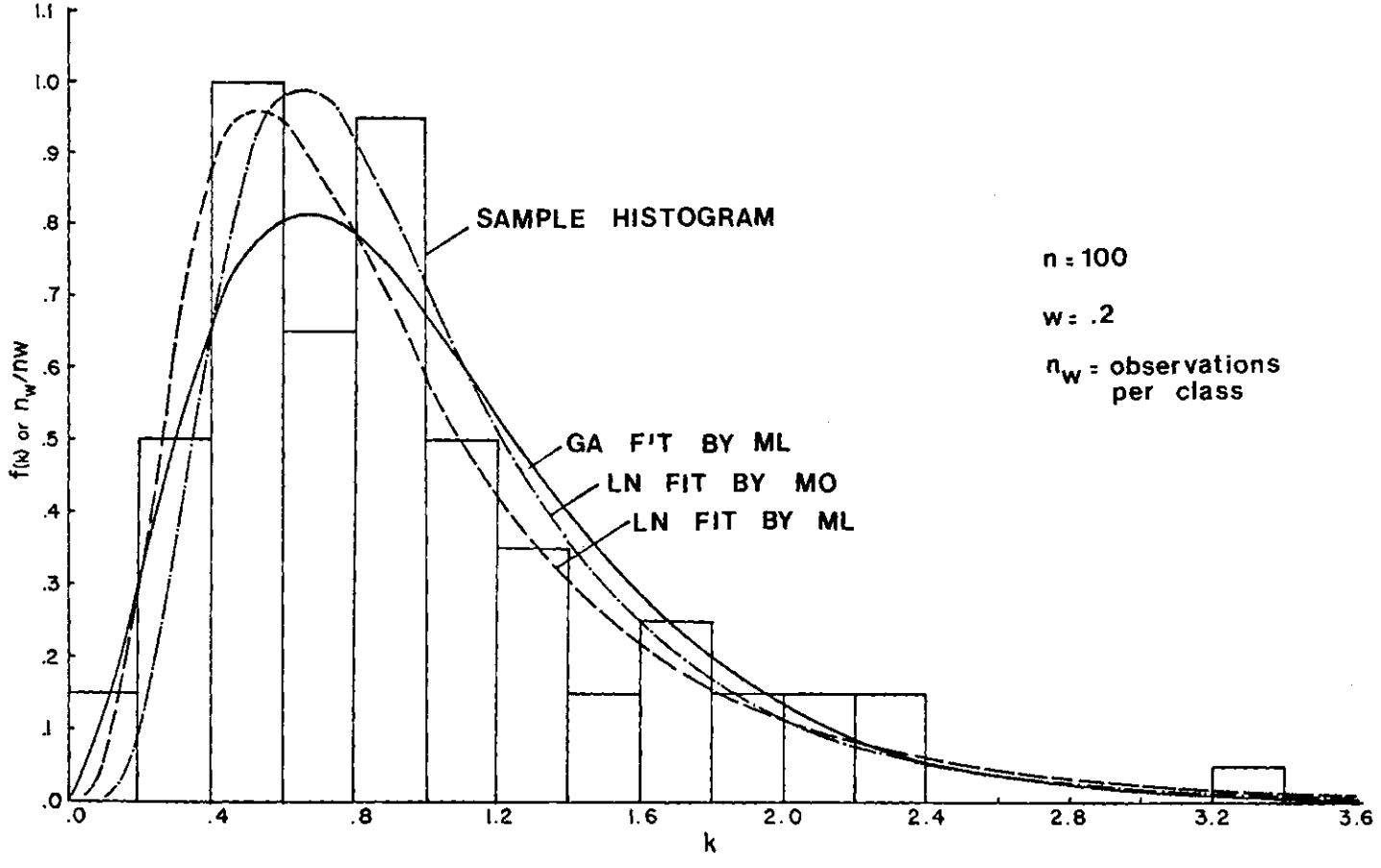


Figure 2.3 Sample 2 from Illustrative Example

CHAPTER III

DIMENSIONLESS FREQUENCY ANALYSIS

Comparison of the statistical properties of hydrologic data at different stations is facilitated by normalizing into dimensionless variables by dividing data items by the sample mean (Markovic, 1964). In this Chapter, dimensionless variables are used in a detailed study of three distributions: the two-parameter lognormal (LN), the two-parameter gamma (GA) and the two-parameter Gumbel (GU) (the Type I asymptotic extreme value distribution of the largest values). The study shows that for all three distributions a single statistical parameter, namely, the variance, affords a means of classifying and visualizing the shapes of the density functions and evaluating percentiles of the distribution.

Let Q_i be the observed hydrologic data at a given station and $\bar{Q}$ be the sample mean. The Q_i may be expressed as dimensionless variables, K_i, by the relation

$$K_i = \frac{Q_i}{\bar{Q}} \tag{3.1}$$

The sample mean, $\bar{K}$, of dimensionless variables K_i given by

$$\bar{K} = \frac{1}{n} \sum_{i=1}^{n} K_i \tag{3.2}$$

will be unity for any sample. Also, it can be easily shown that the variance of K_i, S_k^2, equals C_v^2 , where C_v is the coefficient of variation of the observed data and is given by,

$$C_v = \frac{S}{\overline{Q}} \tag{3.3}$$

in which S = sample standard deviation of Q_i.

Statistical Properties of the Selected Distributions

The three density functions used in this study are given by Table 3.1. Four statistical parameters, namely, the mean (μ_x), the variance (σ_x^2), the skewness coefficient (γ_1) and the kurtosis coefficient (γ_2) are usually all that are useful in providing insight into the characteristics of a frequency distribution. The mean shows the central tendency (or the value about which all other values are clustered). The variance indicates the dispersion or the spread of the values about a central value, and its square root is known as the standard deviation (σ_x). The skewness coefficient (γ_1), which is defined as the ratio of the third central moment (μ_3) to the standard deviation cubed, or

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3} \tag{3.4}$$

describes the asymmetry of a frequency distribution.

Table 3.1. Properties of the Probability Density Functions Selected for Study

Probability Density Function (PDF) (1)	Form of Equation, f(x) (2)	Mean (μ_x) (3)	Variance σ_x^2 (4)	Skewness Coefficient γ_1 (5)	Kurtosis Coefficient γ_2 (6)
Lognormal (LN) Parameters: μ_y and σ_y	$\frac{1}{x\sigma_y\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{\ln x - \mu_y}{\sigma_y}\right]^2}$ $x > 0$, $\ln x = Y$ μ_y = Mean of Y σ_y = Standard Deviation of Y $\eta = \sigma_x / \mu_x$	$e^{\mu_y + \sigma_y^2/2}$	$\mu_x^2 (e^{\sigma_y^2} - 1)$	$\eta^3 + 3\eta$	$\eta^8 + 6\eta^6 + 15\eta^4 + 16\eta^2 + 3$
Gamma (GA) Parameters: C and D	$\frac{C^D x^{D-1} e^{-Cx}}{\Gamma(D)}$ $x > 0$ C = Scale Parameter > 0 D = Shape Parameter > 0	D/C	D/C^2	$2/\sqrt{D}$	$6/D + 3$
Gumbel (GU) Parameters: a and u	$ae^{\left[-a(x-u) - e^{-a(x-u)}\right]}$ $-\infty \le x \le \infty$ a = Dispersion Parameter u = Mode (x = Random Variable)	$u + \frac{.5772}{a}$	$\pi^2/6a^2$	1.1396	4.5

The kurtosis coefficient (γ_2) given by

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} \qquad 3.5$$

in which μ_2 and μ_4 are the population second and the fourth central moments respectively, indicates the flatness of a distribution. If the deviations of variable values from the mean decrease, the fourth central moment tends to zero faster than the square of the second central moment, in which case the kurtosis tends rapidly to zero. This extreme corresponds to a sharp peakedness. If the deviations from the mean increase, the fourth moment tends to infinity faster than the square of the second moment, in which case the kurtosis tends rapidly to infinity. This extreme corresponds to a very pronounced flatness.

Table 3.1 gives the equations of μ_x, σ_x^2, γ_1 and γ_2 for the three distributions selected in terms of distribution parameters.

For dimensionless variables the population mean, μ_k, is unity. By using the relation $\mu_x = \mu_k = 1.0$ in equations given by column 3 of Table 3.1 the following relations may be established between the two parameters for each of the distributions selected.

Lognormal (LN): $\mu_y = -\sigma_y^2/2$ 3.6

Gamma (GA): $C=D$ 3.7

Gumble (GU): $a = \frac{.5772}{1 - u}$ 3.8

Table 3.2. Moments of Dimensionless LN, GA, and GU PDF's

Mean, mu =1.0

Variance	Coefficient of Skewness, γ_1*			Coefficient of Kurtosis, γ_2*		
σ_k^2	LN	GA	GU	LN	GA	GU
0.05	.682	.447	1.140	3.838	3.300	4.5
0.10	.980	.632	"	4.756	3.600	"
(0.1322)[1]	(1.140)	.727	"	(5.394)	3.793	"
0.20	1.431	.894	"	6.850	4.200	"
0.30	1.807	1.095	"	9.320	4.800	"
(0.3247)[2]	1.894	(1.140)	"	9.993	4.948	"
0.40	2.150	1.265	"	12.210	5.400	"
0.50	2.475	1.414	"	15.562	6.000	"
0.60	2.789	1.549	"	19.426	6.600	"
0.70	3.096	1.673	"	23.848	7.200	"
0.80	3.399	1.789	"	28.882	7.800	"
0.90	3.700	1.897	"	34.580	8.400	"
1.00**	4.000		"	41.000		"
1.50	5.511		"	86.062		"
2.00	7.071		"	159.000		"
2.50	8.696		"	269.562		"
3.00	10.392		"	429.000		"

* γ_1 and γ_2 are dimensionless and thus apply to any data.

** The bell shape is not retained by Gamma distribution when $\sigma_k^2 \geq 1.0$

1 The first three moments of LN are equal to the first three moments, respectively, of GU.

2 The first three moments of GA are equal to the first three moments, respectively, of GU.

The Moments and the Shapes of Dimensionless Distributions

The mean (μ) and the variance (σ^2) are sufficient to determine the parameters of LN, GA and GU PDF's since they are two-parameter PDF's. In dimensionless form, the mean μ_k of the PDF's becomes unity and thus the value of the variance, σ^2_k, determines the magnitude of the high moments of the three PDF's. Table 3.2 presents these moments of dimensionless LN, GA and GU PDF's for a wide range of σ^2_k.

Table 3.2 shows that both skewness and kurtosis coefficients rapidly increase with an increase of variance for the lognormal distribution while the increase is not so large for the Gamma distribution. The Gumbel distribution, as a property, has a constant skewness and constant kurtosis. Gamma distribution has, for all values of variance, a lower skewness and lower kurtosis compared to the lognormal distribution. For no value of variance are the skewness and kurtosis coefficients of any one distribution simultaneously equal to those of another distribution. This shows that no two distributions are exactly identical at any value of variance. However, skewness of gamma and Gumbel distributions are equal when the variance is 0.3247 and skewness of lognormal and Gumbel distributions are equal when the variance is 0.1322 (see Table 3.2). At these values of variance the values of kurtosis are not widely different for these pairs of distributions. Hence the LN and GU and the GA and GU may be expected to have similar shapes and approximately equal percentiles when σ^2_k equals 0.1322 and 0.3247, respectively.

To examine how the shapes of frequency distributions change with variance, probability density curves are drawn varying the value of σ_k^2 from .10 to 3.0 for the three PDF's (see Figures 3.1 through 3.3). In these curves the probability density function (PDF), f(k), is shown against the dimensionless variate $k = (Q/\bar{Q})$. Table 3.3* presents the numerical values of variable k (percentiles) for the three distributions at selected values of the cumulative distribution function (CDF), F(k), of each distribution. $\sigma_k^2 = 0.1322$ is the case for which the first three moments of lognormal are numerically equal to the first three moments, respectively, of the Gumbel distribution, and thus the two distributions are practically indistinguishable (see Figure 3.4). Similarly, at $\sigma_k^2 = 0.3247$ the first three moments of the gamma distribution are numerically equal to the first three moments, respectively, of the Gumbel distribution. Thus the gamma and Gumbel distributions become practically indistinguishable at $\sigma_k^2 = 0.3247$ (see Figure 3.5), although there is a slight shift in the modes which occurs since the lower limit of the Gumbel distribution is at negative infinity.

Based on Figures 3.1 through 3.7 and the results presented in Table 3.3, the characteristic features of lognormal, gamma, and the Gumbel distributions are summarized in the following paragraphs.

Characteristic Shapes

The shapes of the distributions can be characterized as follows:

---

* In Table 3.3 the percentiles for gamma distribution were evaluated by using the sub-program PIGAMA from math-science library of CDC CYBER 74 computer. Evaluation of percentiles for lognormal and the Gumbel distributions is easily programmable.

Table 3.3. Numerical Values of Variable K at Selected Frequencies

		*** RETURN PERIOD IN YEARS ***												
		1.01	1.05	1.11	1.25	2.00	5.00	10.00	25.00	50.00	100.00	200.00	500.00	1000.00
							- C D F -							
VAR	PDF	.010	.050	.100	.200	.500	.800	.900	.960	.980	.990	.995	.998	.999
.0500	LN	.584	.679	.735	.810	.976	1.175	1.295	1.437	1.536	1.631	1.724	1.843	1.931
	GA	.554	.663	.726	.809	.983	1.182	1.295	1.424	1.511	1.592	1.669	1.765	1.835
	GU	.633	.708	.754	.816	.963	1.161	1.292	1.457	1.580	1.701	1.823	1.983	2.104
.1000	LN	.465	.574	.642	.735	.953	1.236	1.416	1.637	1.797	1.955	2.112	2.319	2.476
	GA	.413	.543	.622	.729	.967	1.252	1.421	1.616	1.751	1.878	2.000	2.154	2.266
	GU	.481	.567	.652	.740	.948	1.228	1.413	1.646	1.820	1.992	2.163	2.390	2.561
.1322	LN	.414	.526	.598	.699	.940	1.264	1.476	1.742	1.938	2.133	2.329	2.591	2.792
	GA	.351	.486	.571	.688	.956	1.286	1.485	1.718	1.880	2.033	2.181	2.368	2.505
	GU	.403	.525	.600	.701	.940	1.262	1.474	1.743	1.943	2.140	2.338	2.598	2.795
.2000	LN	.338	.452	.528	.637	.913	1.308	1.578	1.928	2.194	2.465	2.742	3.120	3.412
	GA	.256	.394	.487	.618	.934	1.344	1.599	1.902	2.116	2.321	2.519	2.772	2.958
	GU	.266	.416	.508	.633	.927	1.322	1.583	1.914	2.159	2.403	2.645	2.965	3.207
.3000	LN	.266	.378	.455	.570	.877	1.350	1.691	2.150	2.511	2.888	3.281	3.831	4.268
	GA	.167	.298	.393	.535	.902	1.409	1.734	2.131	2.415	2.689	2.956	3.300	3.556
	GU	.101	.285	.397	.550	.910	1.394	1.715	2.119	2.420	2.718	3.015	3.407	3.703
.3247	LN	.253	.363	.440	.556	.869	1.358	1.714	2.198	2.592	2.983	3.405	3.997	4.470
	GA	.151	.279	.374	.518	.894	1.422	1.764	2.182	2.482	2.773	3.056	3.422	3.694
	GU	.065	.256	.373	.532	.906	1.410	1.743	2.165	2.477	2.787	3.096	3.504	3.812
.4000	LN	.219	.326	.402	.519	.845	1.377	1.777	2.333	2.782	3.258	3.765	4.487	5.075
	GA	.111	.229	.322	.469	.870	1.458	1.847	2.329	2.678	3.017	3.350	3.781	4.099
	GU	-.038	.174	.304	.481	.896	1.455	1.825	2.293	2.640	2.984	3.327	3.779	4.122
.5000	LN	.186	.286	.361	.478	.816	1.395	1.846	2.489	3.019	3.592	4.210	5.102	5.841
	GA	.074	.178	.266	.412	.839	1.497	1.945	2.506	2.917	3.319	3.715	4.230	4.617
	GU	-.160	.077	.222	.419	.884	1.509	1.922	2.445	2.833	3.218	3.602	4.108	4.490

Table 3.3. Numerical Values of Variable K at Selected Frequencies - Continued

		*** RETURN PERIOD IN YEARS ***												
		1.01	1.05	1.11	1.25	2.00	5.00	10.00	25.00	50.00	100.00	200.00	500.00	1000.00
							- C D F -							
VAR	PDF	.010	.050	.100	.200	.500	.800	.900	.960	.980	.990	.995	.998	.999
.6000	LN	.160	.256	.328	.444	.791	1.408	1.903	2.625	3.231	3.896	4.622	5.684	6.576
	GA	.050	.138	.220	.364	.809	1.529	2.031	2.669	3.139	3.602	4.059	4.655	5.106
	GU	-.271	-.011	.148	.369	.873	1.557	2.011	2.583	3.008	3.430	3.850	4.404	4.823
.7000	LN	.141	.231	.302	.415	.767	1.416	1.951	2.745	3.423	4.176	5.008	6.237	7.284
	GA	.034	.108	.183	.322	.779	1.555	2.109	2.819	3.347	3.869	4.387	5.066	5.576
	GU	-.373	-.092	.079	.313	.863	1.602	2.091	2.710	3.169	3.624	4.078	4.677	5.129
.8000	LN	.125	.211	.279	.391	.745	1.421	1.991	2.853	3.599	4.435	5.370	6.765	7.966
	GA	.022	.084	.152	.285	.750	1.577	2.179	2.960	3.544	4.124	4.701	5.461	6.031
	GU	-.468	-.168	.016	.266	.853	1.644	2.167	2.828	3.319	3.806	4.291	4.931	5.414
.9000	LN	.113	.194	.260	.370	.725	1.424	2.025	2.950	3.760	4.677	5.713	7.271	8.625
	GA	.015	.066	.127	.252	.721	1.595	2.243	3.093	3.732	4.369	5.005	5.843	6.473
	GU	-.557	-.239	-.044	.221	.844	1.683	2.238	2.939	3.459	3.976	4.490	5.169	5.682
1.0000	LN	.102	.180	.243	.351	.707	1.425	2.055	3.037	3.909	4.904	6.037	7.756	9.263
	GA	.010	.051	.105	.223	.693	1.609	2.303	3.219	3.912	4.605	5.298	6.214	6.904
	GU	-.641	-.306	-.100	.179	.836	1.719	2.305	3.044	3.592	4.137	4.679	5.395	5.936
1.5000	LN	.068	.131	.185	.283	.632	1.415	2.157	3.379	4.517	5.862	7.445	9.943	12.179
	GA	.001	.014	.041	.121	.565	1.646	2.539	3.767	4.718	5.682	6.655	7.954	8.936
	GU	-1.010	-.599	-.348	-.006	.799	1.881	2.598	3.503	4.175	4.842	5.506	6.382	7.045
2.0000	LN	.050	.103	.151	.239	.577	1.395	2.212	3.617	4.970	6.612	8.589	11.773	14.723
	GA	.000	.004	.016	.064	.455	1.642	2.706	4.218	5.412	6.635	7.879	9.549	10.821
	GU	-1.320	-.846	-.556	-.161	.768	2.017	2.845	3.890	4.666	5.436	6.203	7.215	7.980
2.5000	LN	.040	.085	.127	.208	.535	1.371	2.243	3.793	5.324	7.223	9.551	13.376	16.981
	GA	.000	.001	.006	.033	.363	1.614	2.825	4.601	6.027	7.500	9.008	11.043	12.600
	GU	-1.594	-1.064	-.740	-.298	.740	2.138	3.063	4.232	5.099	5.960	6.817	7.949	8.804
3.0000	LN	.032	.072	.111	.186	.500	1.347	2.261	3.928	5.612	7.735	10.378	14.791	19.011
	GA	-.000	.000	.002	.017	.287	1.569	2.909	4.931	6.581	8.295	10.065	12.459	14.298
	GU	-1.842	-1.261	-.906	-.422	.715	2.246	3.260	4.540	5.490	6.433	7.372	8.612	9.549

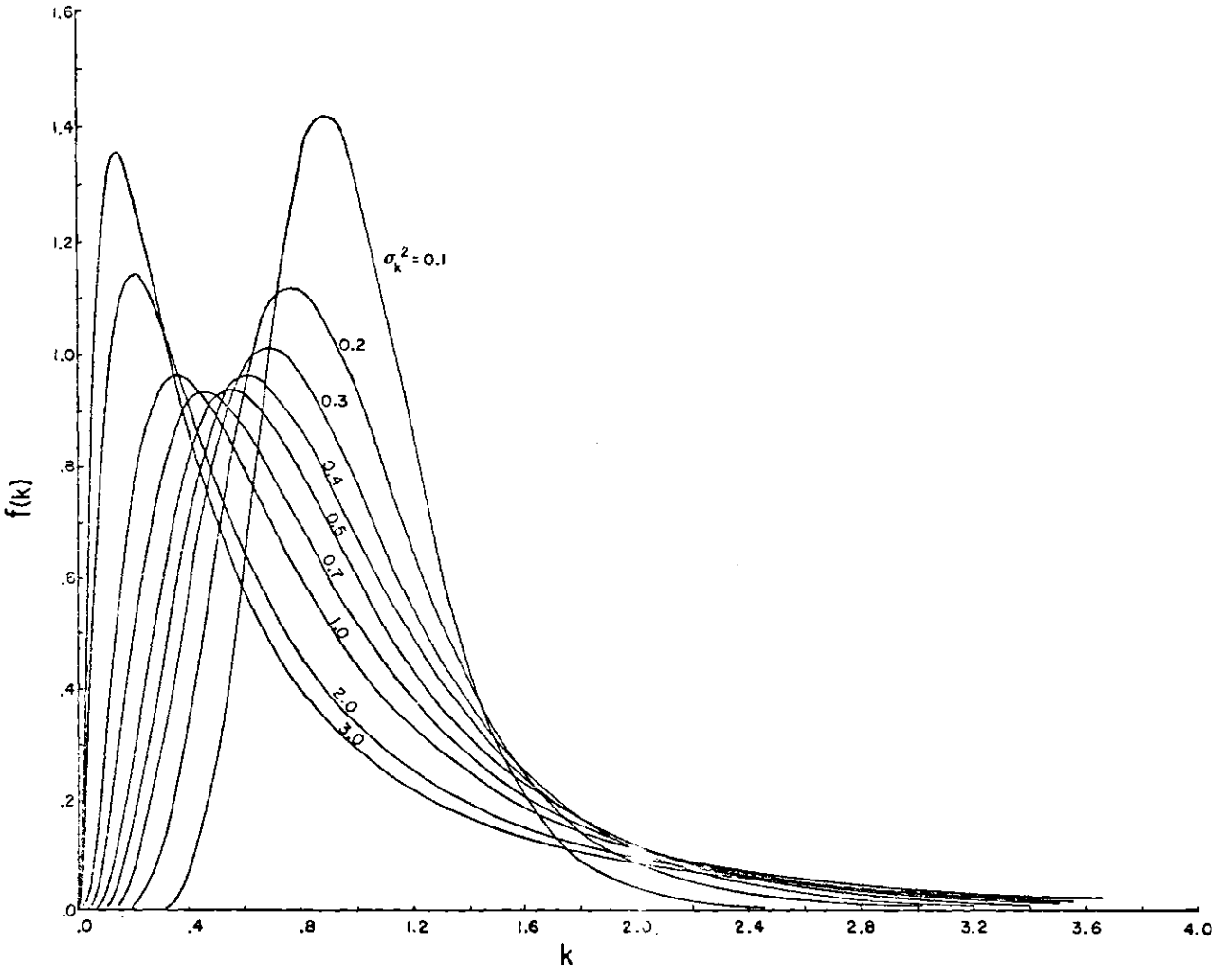


Figure 3.1 Lognormal Densities ($\mu_k = 1.0$)

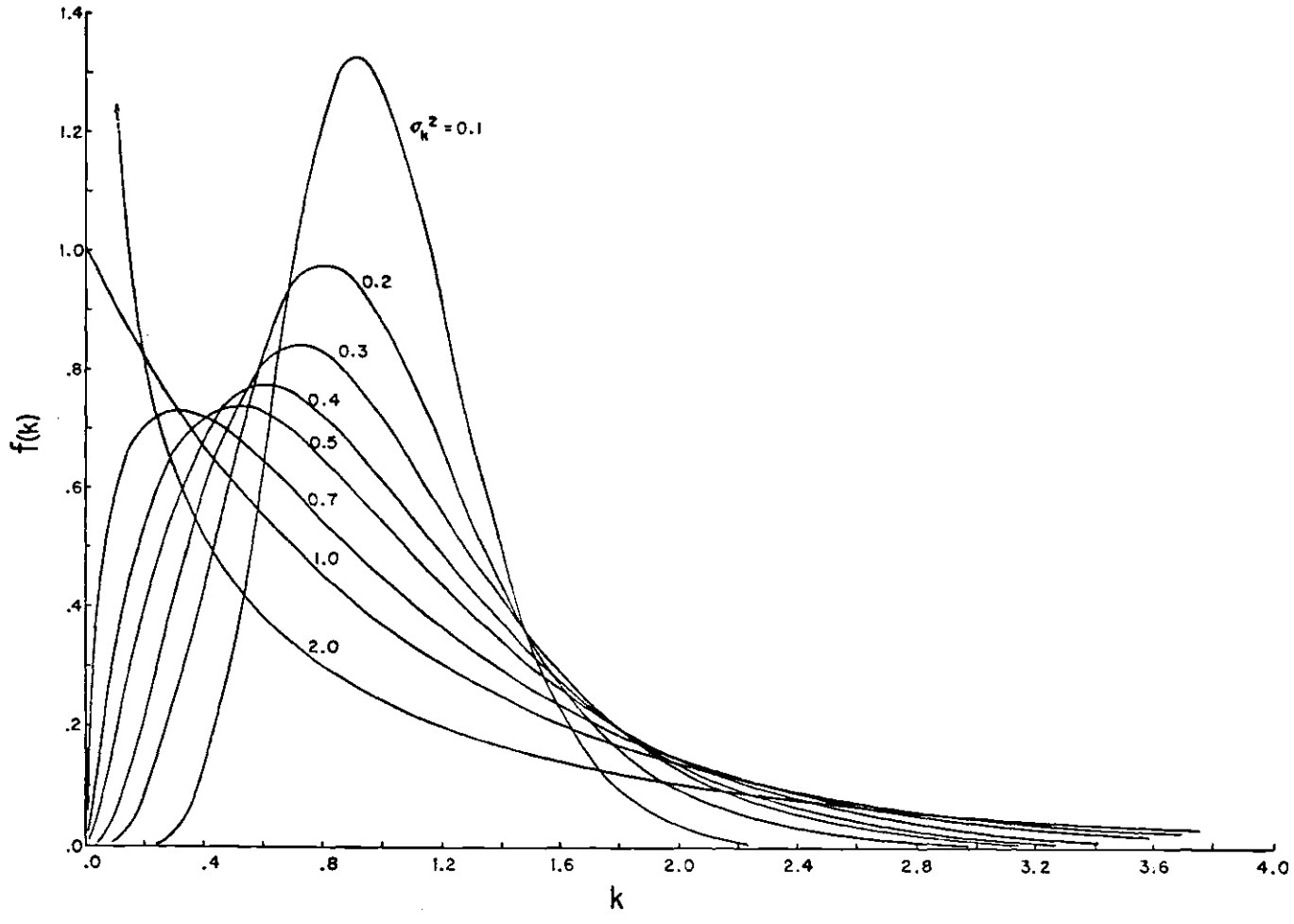


Figure 3.2 Gamma Densities ($\mu_k = 1.0$)

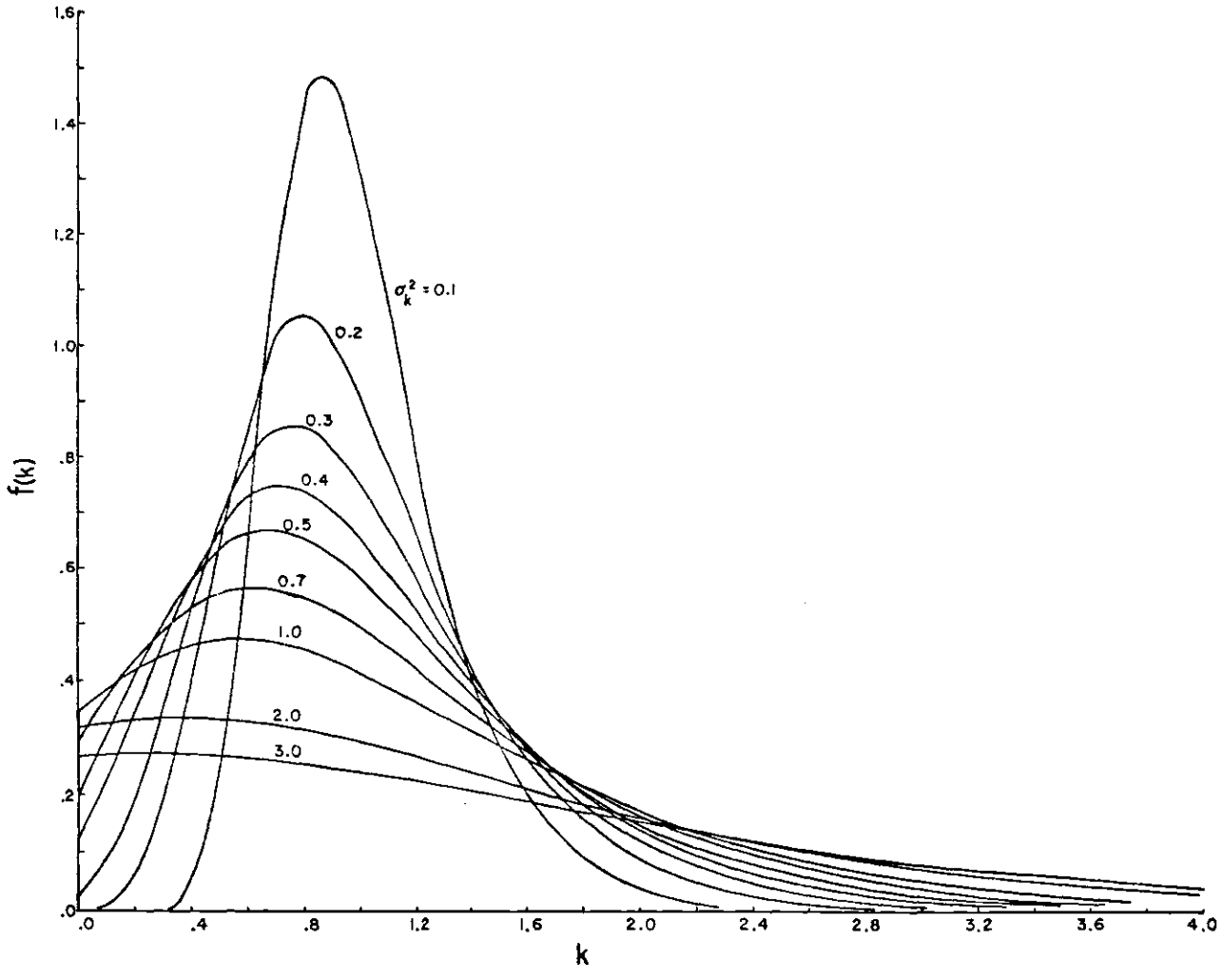


Figure 3.3 Gumbel Densities (μ_k = 1.0)

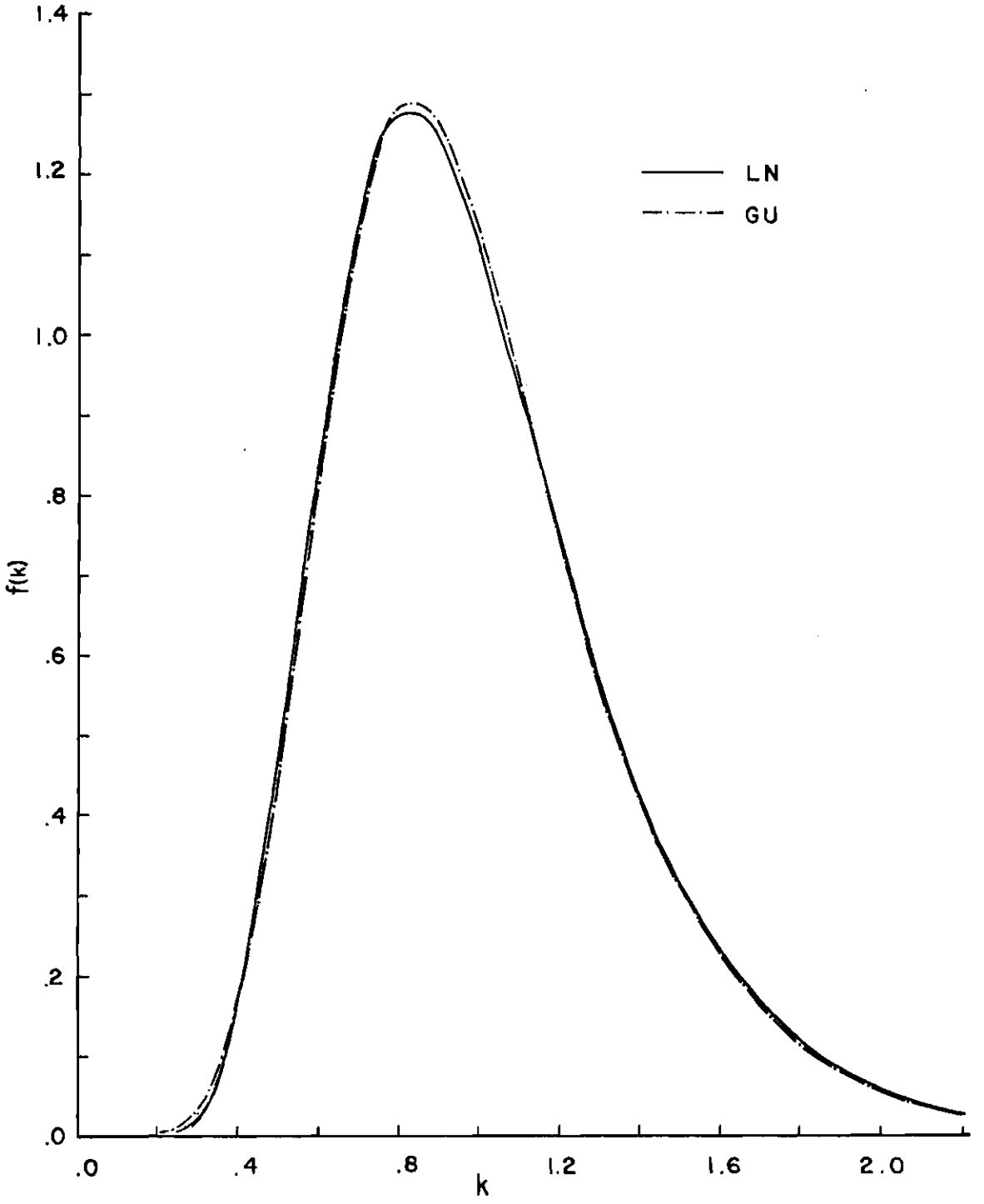


Figure 3.4 LN and GU Densities, $\sigma_k^2 = 0.1322$

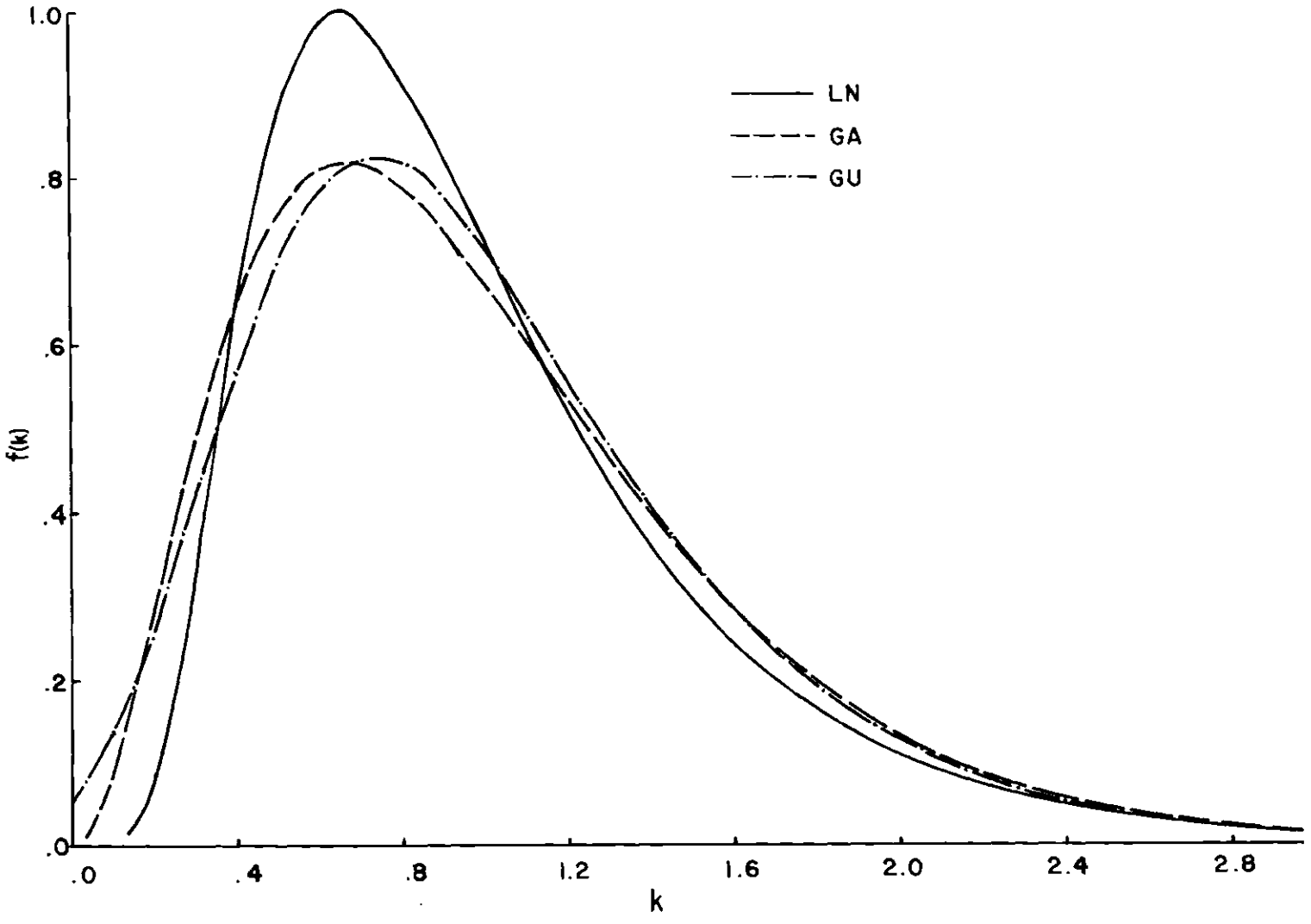


Figure 3.5 LN, GA, and GU Densities, $\sigma_k^2 = 0.3247$

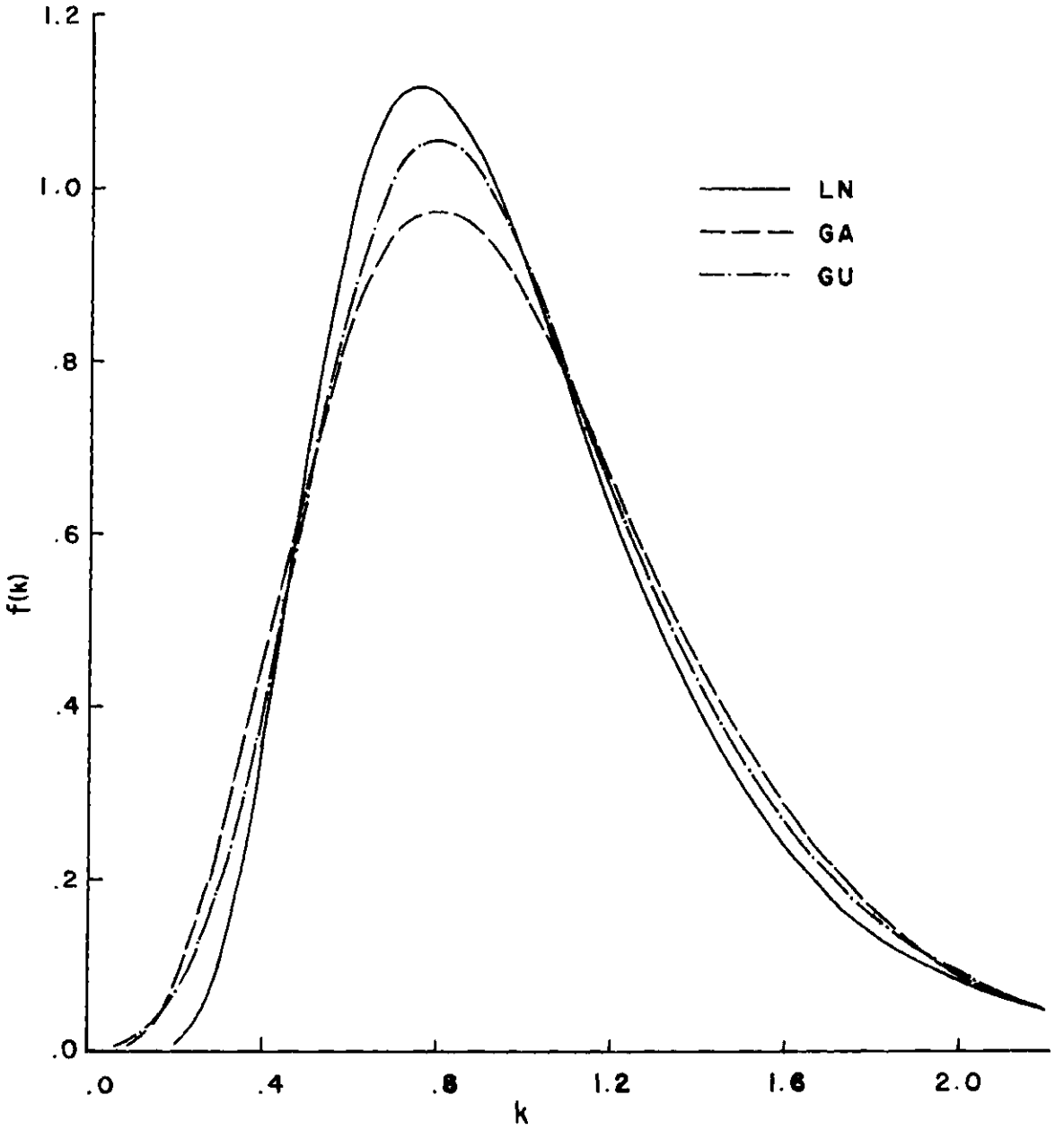


Figure 3.6 LN, GA, and GU Densities, $\sigma_k^2 = 0.20$

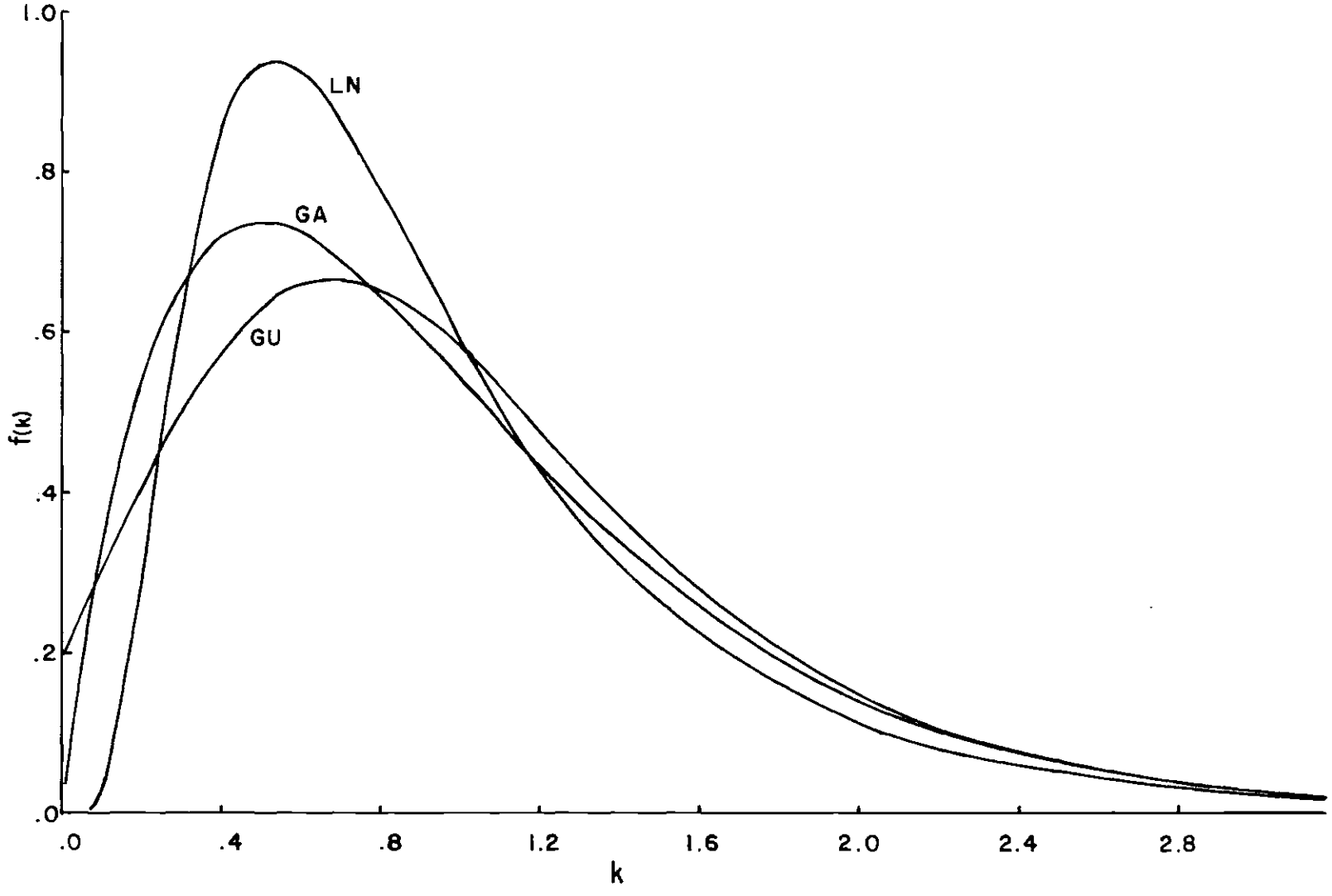


Figure 3.7 LN, GA, and GU densities, $\sigma_k^2 = 0.50$

1. The modes of the three distributions shift towards the origin as σ_k^2 increases (see Figures 3.1 to 3.3).
2. At lower values of σ_k^2 (i.e., up to $\sigma_k^2 = 0.2$), the characteristic shapes of the three density functions are similar except for differences in the probability density at mode (see Figure 3.6).
3. At $\sigma_k^2 = 0.1322$ the lognormal and the Gumbel distributions are practically indistinguishable (see Figure 3.4).
4. At $\sigma_k^2 = 0.3$ to 0.4 the lognormal distribution becomes significantly different from the gamma and Gumbel distributions. In this range of σ_k^2, the gamma and the Gumbel distributions resemble each other closely and become quite similar at $\sigma_k^2 = 0.3247$. The tails of GA and GU are thicker and the probability density at modes lower compared to the LN (see Figure 3.5).
5. At $\sigma_k^2 \geq 0.5$ all three distributions become markedly different from each other even at the same mean and the same variance (see Figures 3.7 and 3.1 through 3.3). The large differences between skewness coefficients and between kurtosis coefficients (see Table 3.2) are indicative of the differences in the characteristic shapes of the three distributions when $\sigma_k^2 \geq 0.5$.

Gamma Distribution

The gamma distribution has, in general, a thicker lower tail, i.e., the curve rises more steeply near the origin and the mode generally

occurs at a lower value of k, compared to the LN and as σ_k^2 increases from 0.4 its lower tail grows thicker and thicker and the mode shifts towards the origin. The gamma becomes an exponential distribution at $\sigma_k^2 = 1.0$ (see Figure 3.2). Thus, the gamma distribution does not retain a bell shape when $\sigma_k^2 \geq 1.0$ and may not be suitable for many important classes of data such as flood peaks and storm rainfall. (However, for a hydrologic sample with many values near zero and $S_k^2 \geq 1.0$ GA may be applied). The thicker lower tail of GA in the range of $\sigma_k^2 = 0.4$ to about 1.0 indicates its particular adaptability to samples which have quite a few low valued variates in this range of σ_k^2.

Gumbel Distribution

Since the lower limit of its random variable k is $-\infty$, a portion of the Gumbel density curve always extends into the negative range of k. However, a hydrologist need not be concerned about this fact as long as the area of the GU density curve in the negative range of k is insignificant. Table 3.4 presents the values of CDF, F(k), at k = 0.0 for different values of σ_k^2. These CDF values represent the proportion of the total area of the density curve in the negative range of k at each value of σ_k^2. Table 3.4 shows that the area of Gumbel curve in the negative range of k becomes more and more significant as σ_k^2 increases above a value of 0.5. Since negative data do not exist in hydrology, (unless made negative by some transform) use of the Gumbel distribution in frequency analyses for samples with such variance may not be meaningful. Figure 3.3, in fact, shows the Gumbel density curves with the negative side truncated. If a 5% limit is set on permissible

Table 3.4. Gumbel distribution - CDF at k=o

Variance σ_k^2 $(=n_Q^2)$	F (k=0)* = portion of total area under the curve lying in the negative range of k.
.05	.0
.10	.0
.20	$.51 \times 10^{-4}$
.30	$.29 \times 10^{-2}$
.4	.014
.5	.032
.6	.053
.7	.074
.8	.095
.9	.114
1.0	.132
2.0	.249
3.0	.308
5.0	.369
10.0	.430
∞	.570

* $F(k=0) = e^{-e^{au}}$, and $au = \left(\frac{\pi}{\sigma_K\sqrt{6}} - 0.5772\right)$

negative area of GU curve for its applicability, it is not desirable to use a GU distribution for data having a S_k^2 larger than 0.6, and with a 10% limit one might use it for data samples having its S_k^2 up to about 0.8 to 0.9. At σ_k^2 = 0.5 to 0.9 GU distribution has a distinctly different characteristic shape. It is a relatively flat distribution compared to LN and GA.

Lognormal Distribution

Of the three distributions LN has a larger skewness at $\sigma_k^2 \geq 0.2$ (see Table 3.2). As a result the value of the PDF at the mode is larger and it has thinner rising and falling limbs compared to GA and GU for a wide range of σ_k^2 (see Figures 3.6 and 3.7). In the foregoing paragraphs, it has been pointed out that both GA and GU distributions become unsuitable for application in many instances in hydrology due to their inherent characteristics when σ_k^2 has a value of about 1.0 and above. Also, the high skewness of lognormal distributions at such high variances causes a large proportion of area under the probability density curve to shift towards the origin (for example, the CDF = 0.5 when the dimensionless variable K = 0.71, 0.58 and 0.5 for σ_k^2 = 1.0, 2.0 and 3.0, respectively). This indicates that for a lognormal distribution to be applicable to a set of data having high variance it is also required that the data contain a high proportion of low valued variables. This would require that for data sets having their S_k^2 = 1.0, 2.0 and 3.0 about 50% of the variates should have their values less than 0.7, 0.6 and 0.5 of the mean, respectively, for satisfactory application of the lognormal distribution. If this condition is fulfilled, lognormal

is the only one of the three distributions which appears to be usable when S_k^2 of the given data is 1.0 or above.

Graphical Comparison

The probability density curves shown by Figures 3.1 through 3.3 may also be plotted as probability distribution curves on log-probability paper using data given in Table 3.3 for different values of σ_k^2. Figure 3.8 shows LN, GA and GU densities as distribution curves on log-probability paper for σ_k^2 = 0.1, 0.5 and 1.0, respectively. (It may be noted that hydrologists often plot their data on log-probability paper and compare how well their data fit a selected distribution). Figure 3.8 shows that, when plotted on a log-probability paper, the three distributions appear to differ only slightly and then at the tails when the variance is low (Figure 3.8a, σ_k^2 = .1) and show marked differences at high variance (Figure 3.8b, σ_k^2 = 1.0). Even though Figure 3.7 shows considerable difference between GA and GU density curves at σ_k^2 = 0.5, they are almost indistinguishable on Figure 3.8c except at the lower tail (the lower tail of GU distribution may not be expected to compare well since k extends to $-\infty$). Thus, a visual comparison may not be adequate to distinguish the differences on a log-probability plot.

Percentiles (Predictions for Various Return Periods)

One of the main uses hydrologists make of probability density functions is to predict the magnitude of some hydrologic variable associated with some exceedence probability e.g., the 100-year flood

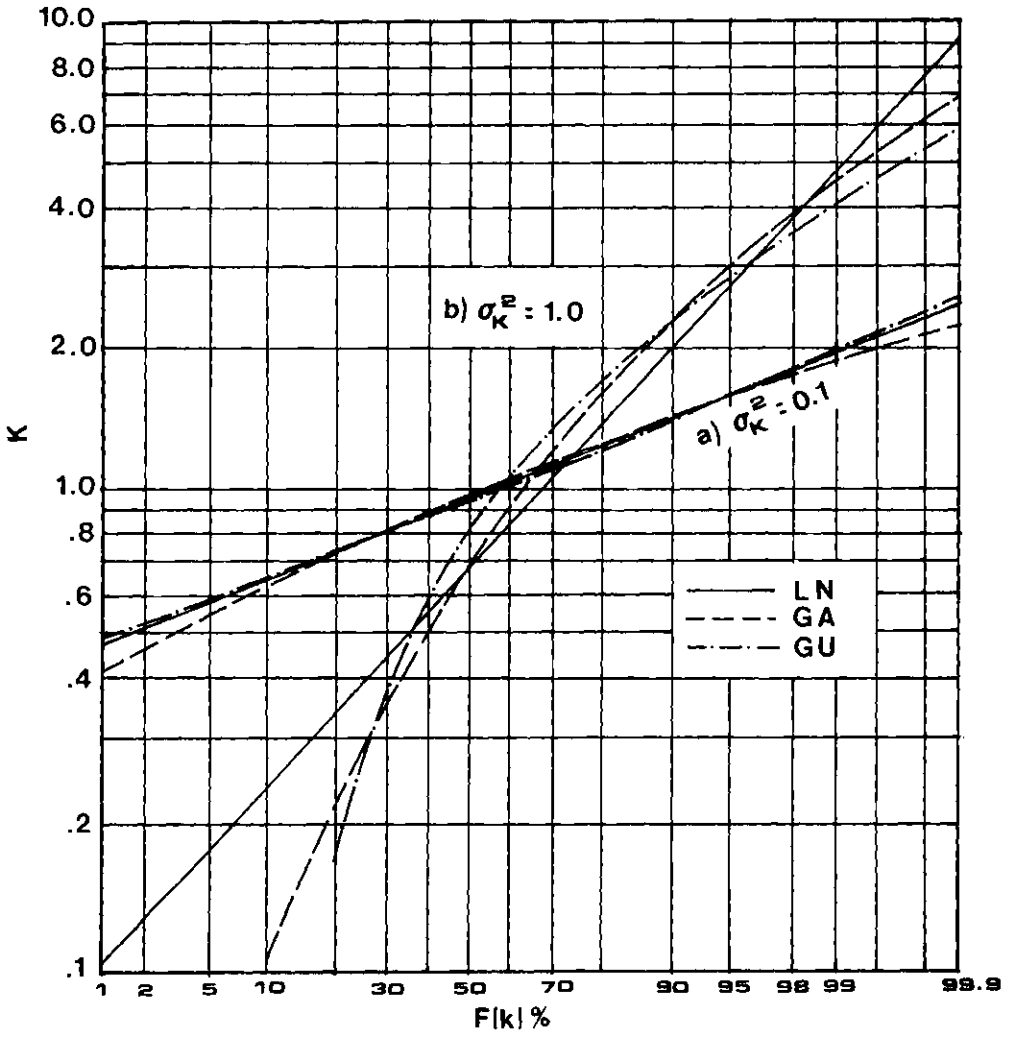


Figure 3.8 Log-probability plots of LN, GA, and Gumbel Distribution; a) $\sigma_k^2 = 0.10$, b) $\sigma_k^2 = 1.0$

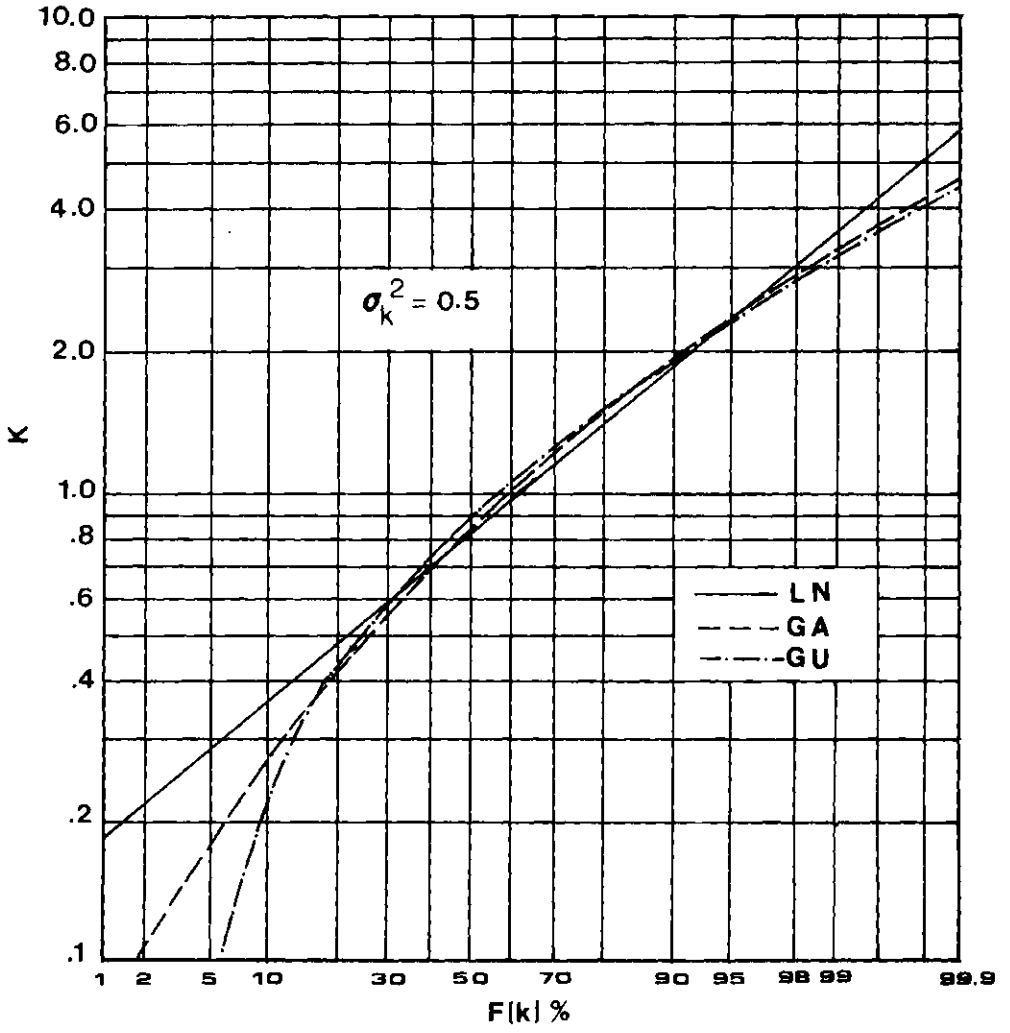


Figure 3.8 Log-probability plots of LN, GA, and Gumbel Distribution; c) $\sigma_k^2 = 0.5$

flow. Table 3.3 presents selected percentiles of LN, GA and GU distributions for $\sigma_k^{2'}$s ranging from 0.05 to 3.00. To portray how the predictions given by the three distributions vary with σ_k^2, the data showing how K_t varies with σ_k^2 presented in Table 3.3 may be plotted as shown in Figure 3.9. The curves show the patterns in the predictions given by the three distributions. The differences among the predictions given by the three distributions at a given variance change with the return periods (t). When t = 2 years (i.e., CDF = 0.5), K_t decreases with an increase of σ_k^2 (Figure 3.9a) because the increasing positive skewness (see Table 3.2) shifts more and more area of the density curves towards the origin. Predictions given by the lognormal distribution are higher than those given by the gamma and Gumbel distributions when $t \geq 50$ years and σ_k^2 exceeds 0.15, and Gumbel gives the lowest predictions of the three when σ_k^2 is larger than about 0.4. (See Figure 3.9). Lognormal predictions are the lowest of the three for a wide range of σ_k^2 for $t \leq 25$ years.

Table 3.5 presents the percentages by which the LN predictions exceed the GA and GU predictions for return periods of 50, 100, 200 and 1000 years as σ_k^2 varies from 0.2 to 3.0.

This chapter provides quantitative information on how the selected probability density function affects frequency analysis. As the variance of the population increases, the shapes of the density functions and the resulting predictions vary more and more among the density functions commonly used. If an appropriate choice of PDF's is not made when the data samples have a S_k^2 larger than 0.2, erroneous predictions will result.

Table 3.5. Percent by Which Lognormal Predictions are Higher Than Gamma and Gumbel Predictions

σ_K^2	Return Period Years							
	50		100		200		1000	
	GA	GU	GA	GU	GA	GU	GA	GU
0.2	3.7	1.6	6.2	2.6	8.8	3.6	15.4	6.5
0.3	4.0	3.8	7.4	6.2	11.0	8.8	20.0	15.3
0.4	3.9	5.4	8.0	9.2	12.4	13.2	23.6	23.1
0.5	3.5	6.6	8.2	11.6	13.3	16.7	26.5	30.1
0.6	2.9	7.4	8.1	13.6	13.8	20.0	28.7	36.3
0.7	2.3	8.0	7.9	15.2	14.1	22.8	30.5	41.9
0.8	1.5	8.4	7.5	16.5	14.2	25.1	32.0	47.1
0.9	0.7	8.7	7.0	17.6	14.1	27.2	33.1	51.7
1.0*	-0.1	8.8	6.5	18.5	13.9	29.0	34.2	56.0
1.5*	-4.3	8.2	3.2	21.1	11.9	35.2	36.3	72.8
2.0*	-8.2	6.5	-0.3	21.6	9.0	38.4	36.1	84.4
2.5*	-11.7	4.4	-3.7	21.2	6.0	40.0	34.8	92.8
3.0*	-14.7	2.2	-6.8	20.2	3.1	40.7	33.0	99.0

* Gamma distribution does not retain its bell shape for $\sigma_K^2 \geq 1.0$

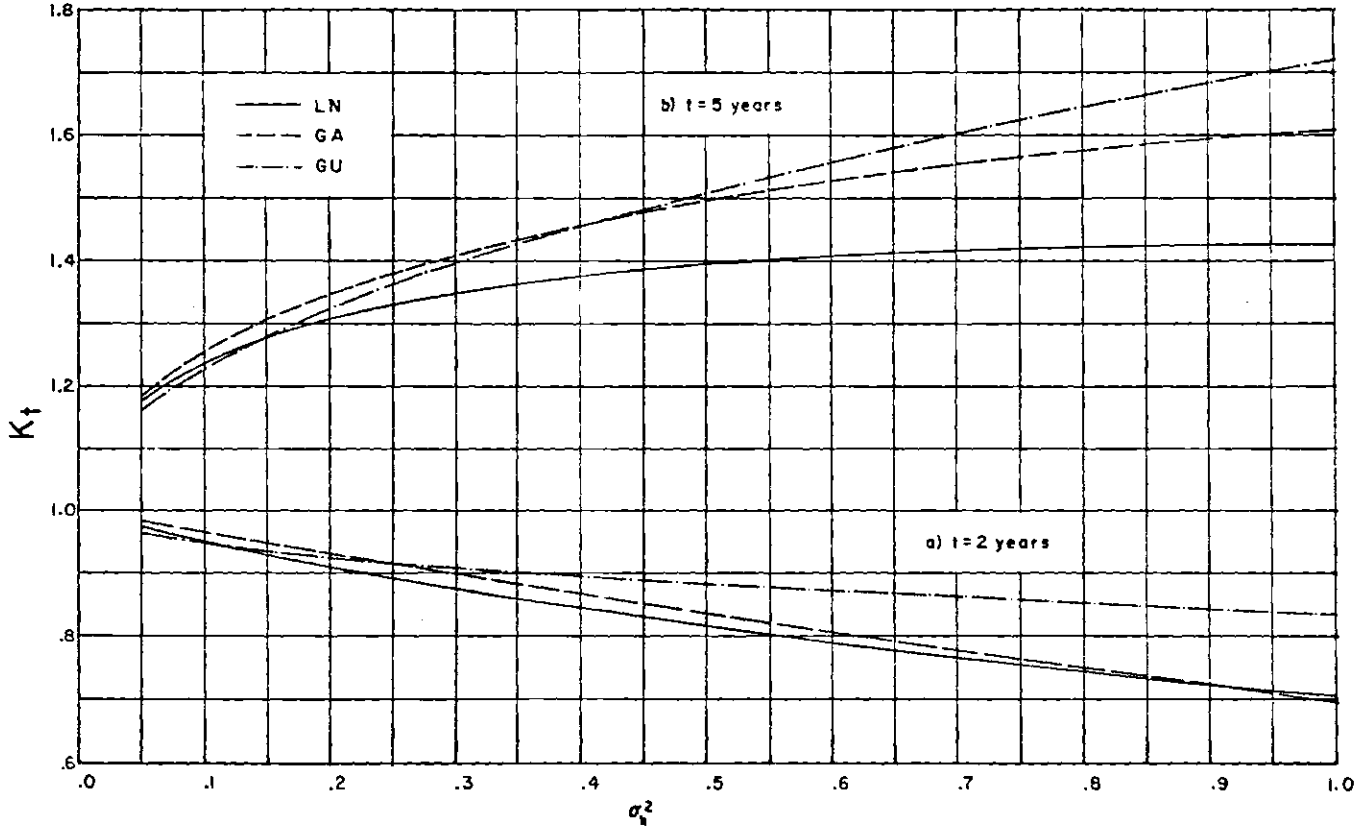


Figure 3.9 K_t versus σ_k^2

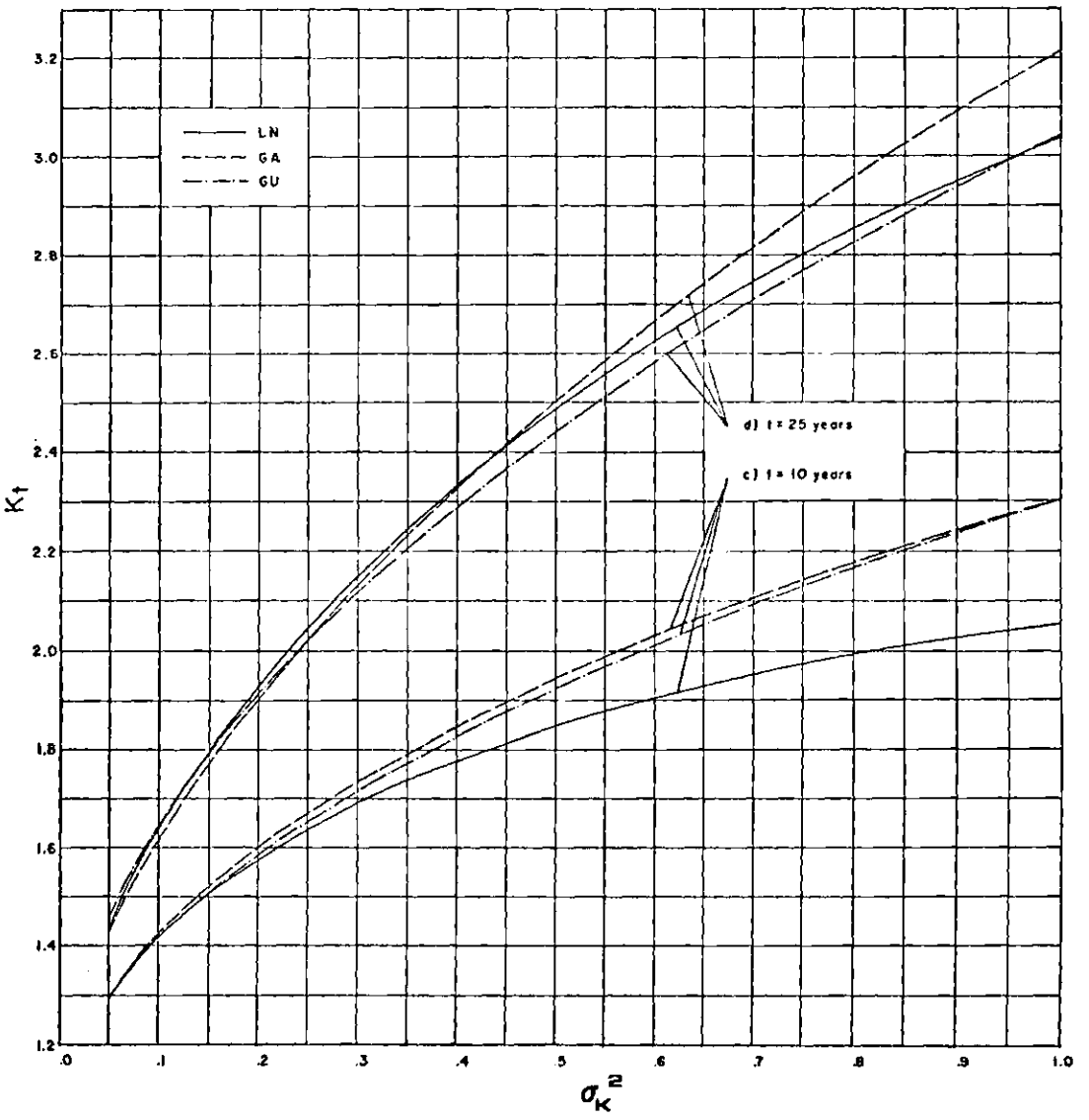


Figure 3.9 K_t versus σ_k^2 (Continued)

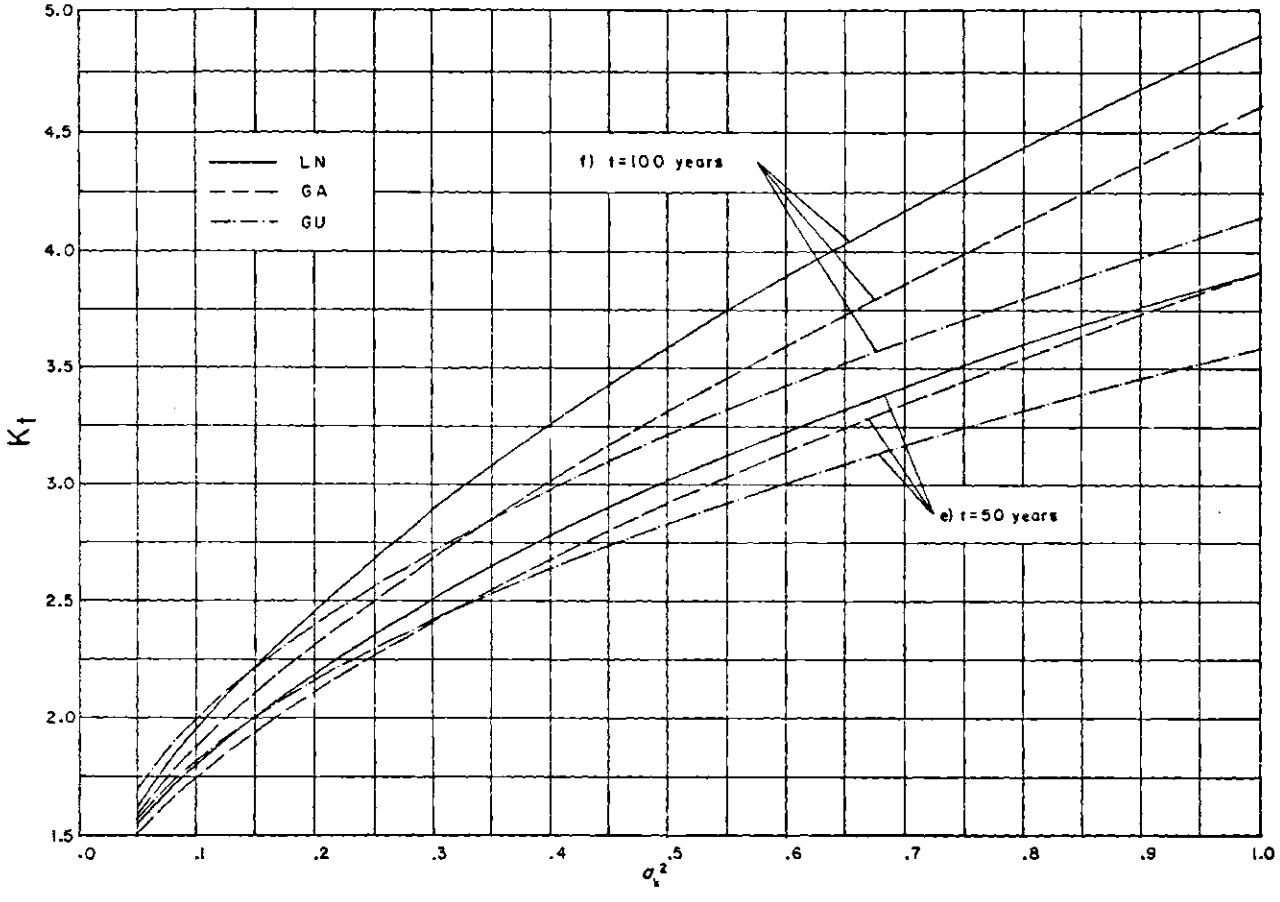


Figure 3.9 K_t versus σ_k^2 (Continued)

CHAPTER IV

SOME CRITERIA TO SELECT THE PROBABILITY DENSITY FUNCTION OF "BEST FIT"

It has been shown in Chapter III that the shapes of density functions vary with the magnitudes of statistical moments. The dimensionless forms of LN, GA and GU distributions (Chapter III) display different shapes at different values of σ_k^2, and the purpose of this chapter is to show how the relationship between the shapes of the density functions and the moments can be used to formulate an analytical criterion to select a probability density function of the "Best Fit". Other criteria which can also help discriminate among the PDF's are also discussed, and a scheme of numerical experiments to verify the validity of the selection "Best Fit" criterion is outlined.

Some symbols used in this chapter and subsequent chapters are as follows:

NO.	SYMBOL	MEANING
1	μ_F	Mean of fitted distribution (i.e., mean based on the estimated parameters, see Table 3.1)
2	σ_F^2	Variance of fitted distribution (i.e., variance based on estimated parameters)
3	K_{S100}	100-year value based on the PDF fitted to the sample
4	LN.GA	LN data are fit to GA PDF. LN.GA may be replaced by any pair of PDF's to give the corresponding meaning

NO.	SYMBOL	MEANING
5	LN.GA/LS	Synthetic LN data are fit to GA by LS. (LN.GA may be replaced by any pair of PDF's and LS may be replaced by MO, ML or MCS to give the corresponding meaning)
6	(LN).GA/LS	Real data judged to be best fit by LN PDF are fit to a GA by LS. ((LN).GA may be replaced by any pair of PDF's and LS may be replaced by MO, ML or MCS to give the corresponding meaning.)

Criterion Based on Simultaneous Fit of Moments and Shape of Sample Distribution

Since the hydrologist generally has a very limited sample of the population of a hydrologic variable, he cannot know the correct probability density function for representing the population with certainty. He must do the best he can by trying to match his selection to known properties of the empirical data. Two properties that seem particularly important for this purpose are the shape of the sample distribution and its moments.

Assume that a sample is taken from a known PDF, f. The moments of the sample will be, on the average, equal to the moments of f and frequencies of the sample events represent, on the average, the shape of f. If f is fit to the sample, the parameter estimates given by MO, ML and LS (see Chapter II for a description of these methods) can be expected to be approximately equal. For example, the MO method, while fitting the moments of f to the sample may well also reproduce the shape of f. Similarly, the ML and LS methods, in addition to fitting the shape of f to the sample distribution, may well also reproduce its

moments.

What will happen if a sample from one PDF is fit to a different PDF by the above three statistical estimation methods? The method of moments will present a fit of the hypothesized PDF with its moments (at least the first m number, where m is number of parameters of PDF) equal to sample moments but with the shape of the sample distribution altogether ignored. On the other hand, the ML and LS methods will present a "fit" which will approximate the shape of sample distribution but have moments that may (greatly) differ from the sample moments. This, in turn, will result in over or under estimates of predicted events when estimates are based on ML or LS methods.

To illustrate, a lognormal synthetic sample of size 100 and S_k^2 = 0.628 was fit to LN by ML and to GA by ML and MO methods, and the shapes of the fitted distributions are plotted with the sample histogram in Figure 4.1. Figure 4.1 shows that σ_F^2, LN.LN/ML is approximately equal to S_k^2 and the LN distribution approximates the sample histogram. (LN.LN/MO also will be practically equal to LN.LN/ML since the variances of the two fits are approximately equal.) The LN.GA/MO fit greatly differs in shape from both the LN fit and the sample histogram, but has its variance equal to S_k^2. On the other hand, the LN.GA/ML fit approximates the shape of the sample distribution; but σ_F^2, LN.GA/ML was found to be only 72% of S_k^2. Predictions by LN.GA/MO would be larger than predictions by LN.GA/ML, but predictions by LN.LN/MO and LN.LN/ML would be approximately equal (values of K_{S100} are shown in Figure 4.1).

Figure 4.2 shows the results of fitting a low variance GA sample

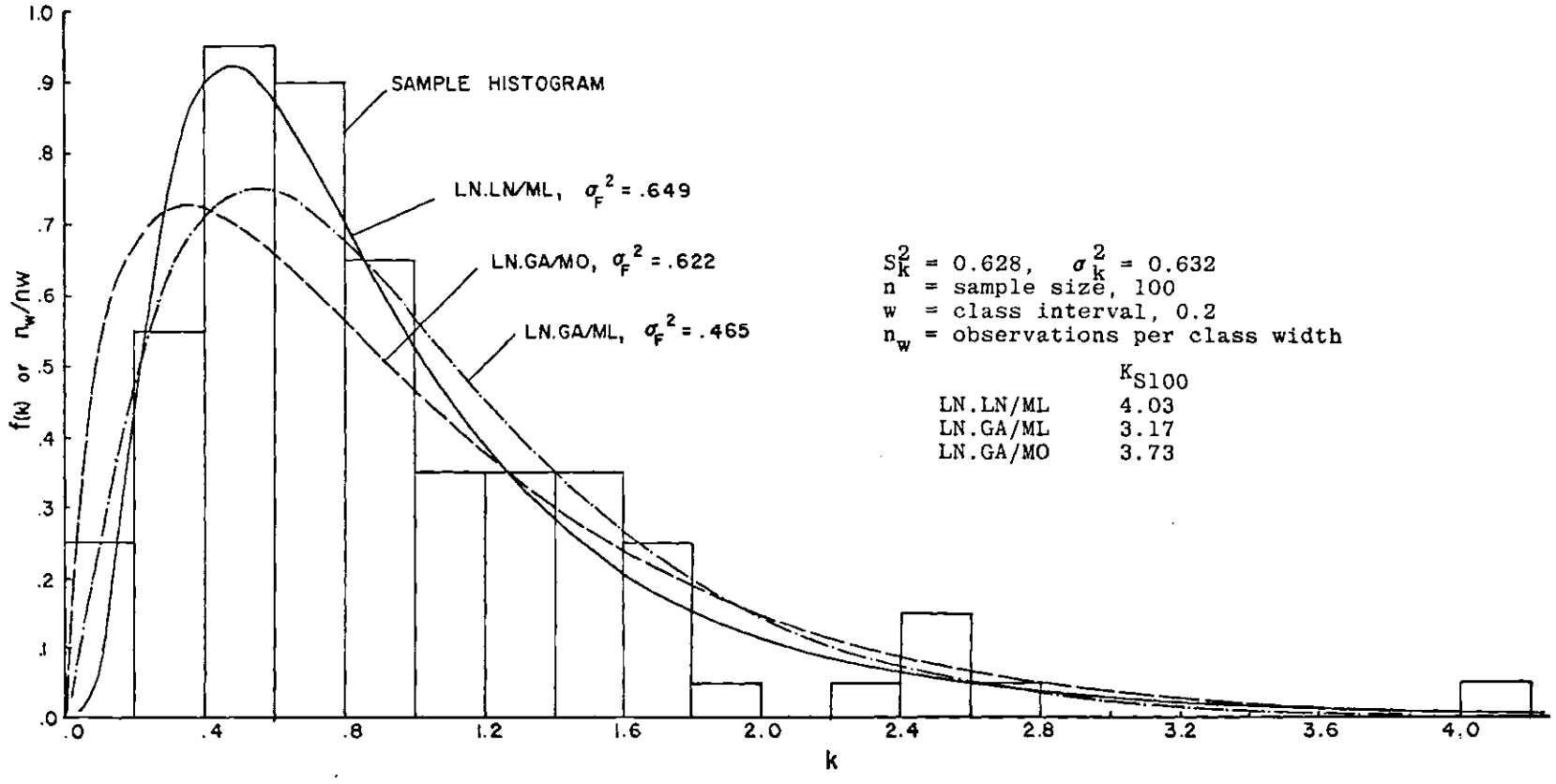


Figure 4.1 LN Sample Fit to LN and GA

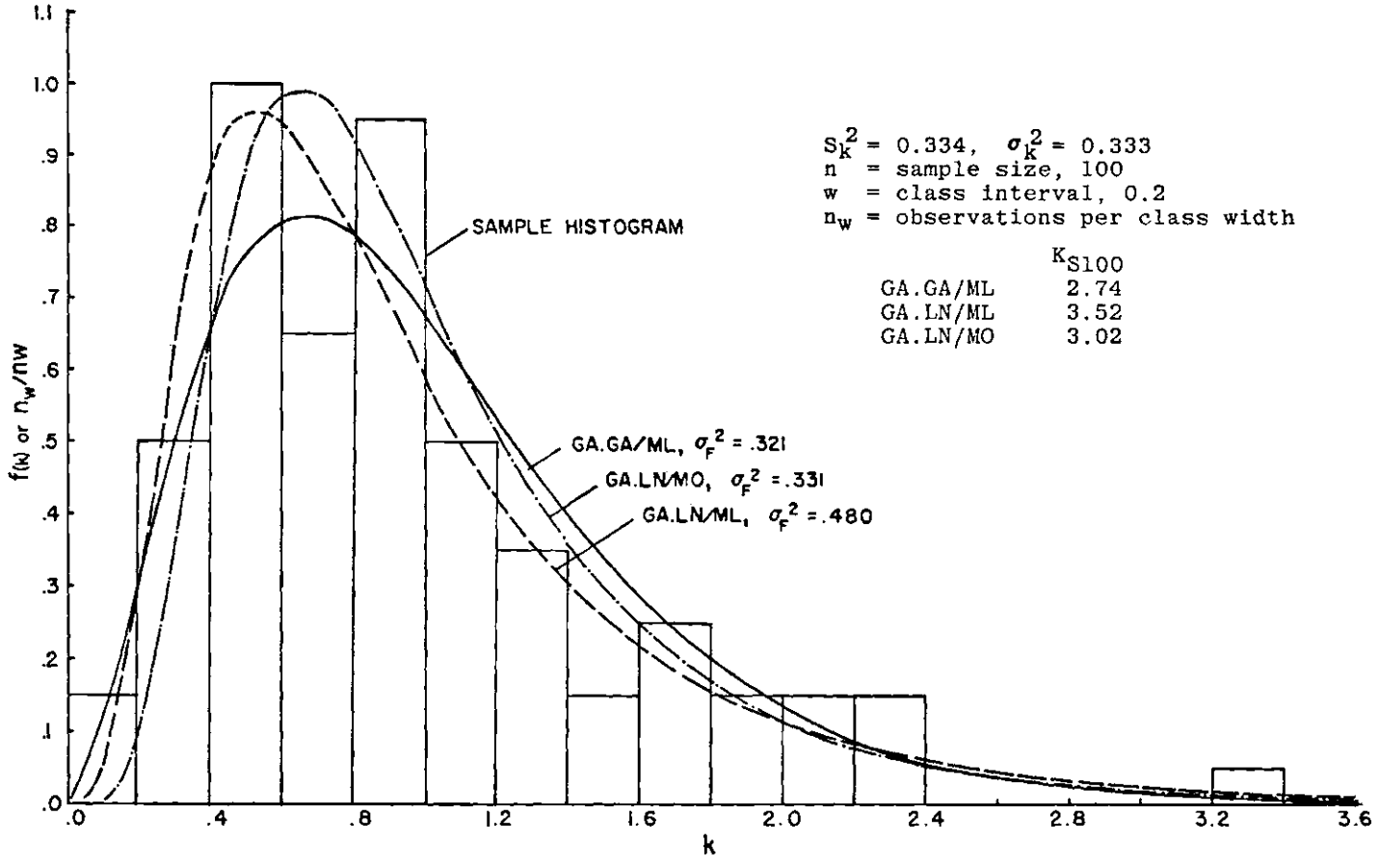


Figure 4.2 GA Sample Fit to GA and LN

(S_k^2 = 0.334) to GA and LN. GA.LN/ML better approximated both the limbs of sample distribution than GA.LN/MO. σ_F^2, GA.LN/ML was found to be 1.44 times S_k^2, and this would result in larger predictions by GA.LN/ML compared to predictions by GA.LN/MO. Figure 4.2 shows the values of K_{S100} by different fits.

One may wonder what will happen when the shape of sample does not conform with shape of parent population. The answer may be found in the GA sample of Figure 4.2. The histogram of the sample shows that the sample, after all, does not well conform with the shape of parent population, which is, in this case, approximately represented by GA.GA/ML density curve. Nevertheless, it was found that σ_F^2, GA.GA/ML $\approx S_k^2$. In general, samples may be regarded as noise-corrupted as shown by Figure 4.3; and, in fact, estimation is defined as extracting information concerning a parameter from noise-corrupted observations (see Chapter II). Some samples may be more noise-corrupted than others. Since the ML method attempts to maximize the probability of joint occurrence of $f(x_i)$ (see Chapter II), it may be regarded that ML method, in general, fits a proposed PDF to the overall shape of the sample although the sample is noise-corrupted. However, anomalous observations may be expected to influence the fit by a shape fitting method. For example, Figures 3.1 and 3.2 show that as σ_k^2 increases the rising limbs of LN and GA density curves shift closer and closer to the origin. Due to the above phenomenon if a sample containing a large number of low valued variables, but having a relatively low sample variance, is fit to LN or GA by ML/LS/MCS it will result in larger σ_F^2 compared to S_k^2. Similarly, if a single

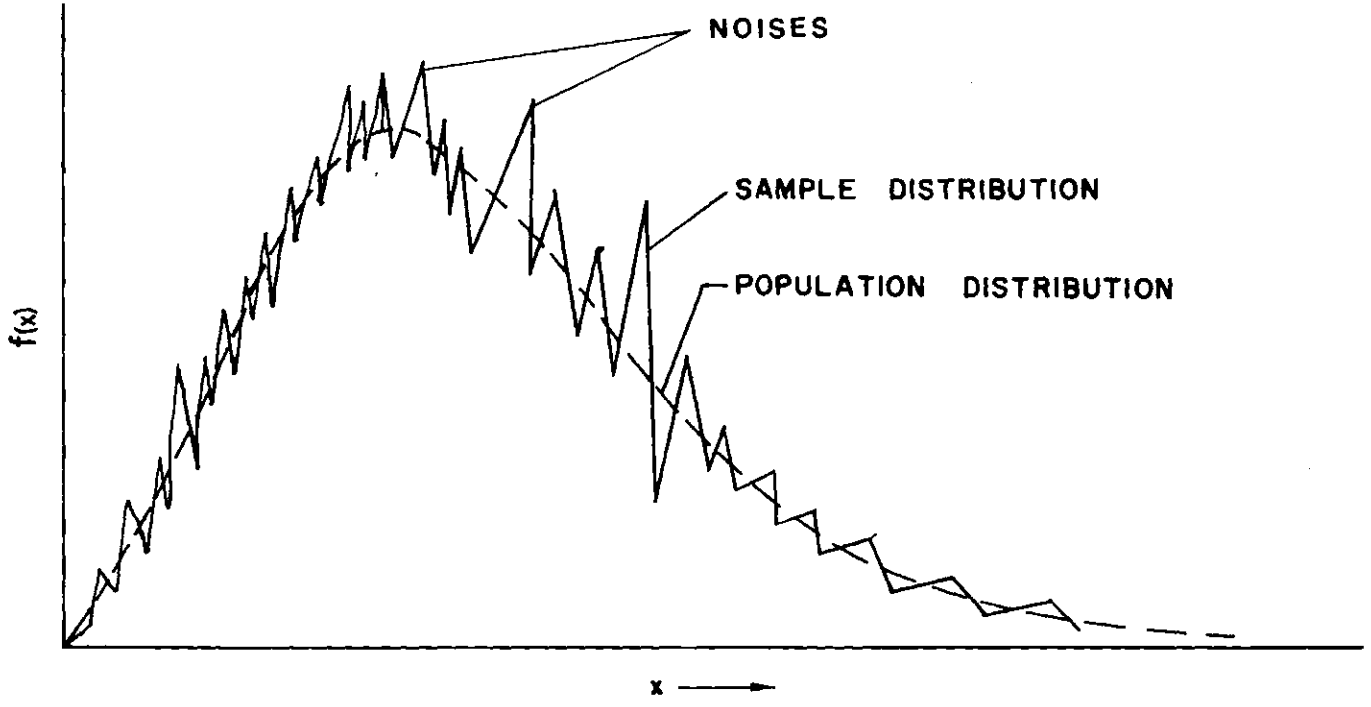


Figure 4.3 Noise Corrupted Sample

observation is far removed from the trend of the other observations lying way in the upper tail (i.e., an outlier) a fit by ML/LS/MCS methods may ignore greatly the presence of such an observation in the sample. A detailed discussion on samples with such anomalous observations is deferred to Chapter VI.

The foregoing discussion indicates that when computations are made by the shape fitting estimation methods like ML, LS or MCS the moments of the fit would be equal to the sample moments only when the hypothesized PDF is approximately the same as the sample distribution. In such case the moments of the sample as well as the distribution of the sample are fit, hence the hypothesized PDF is optimal. The 'Best Fit' criterion based on simultaneous fit of moments and shape of sample distribution may be stated as: The PDF of 'Best Fit' is that PDF whose moments of fit by ML/LS/MCS would be approximately equal to the sample moments.

Criteria Based on Goodness-of-Fit Tests and Least Squares Fit

Hydrologic data such as flood peaks and precipitation have been fit by a variety of probability density functions and, in general, no distribution has been proved to be generally better than the others by statistical tests of goodness-of-fit (see Markovic (1965), Schulz et. al. (1973)). In applying the goodness-of-fit tests to select a PDF, the procedure hitherto has been 'to reject or fail to reject' the hypothesis that the sample is from a hypothesized PDF. This procedure, in general, led to a variety of PDF's being found applicable to the same sample. In this section, parameters of several goodness-of-fit tests are suggested for examination as possible criteria for discriminating between PDF's.

in which c is a critical value for a given significance level. Values of c are furnished in most standard books on statistics (see, for example, Benjamin and Cornell, 1970).

On the basis of Equation 4.3 one may intuitively expect that D_o will be a minimum for a close fit. In fact, D_o will be zero for a perfect fit. When data samples belonging to a particular PDF are fit to various PDF's, including the parent PDF, the values of D_o might be expected to be the smallest, on the average, for the parent fit. Thus, the smallest value of D_o will be examined as a possible criterion to discriminate PDF's by K-S test.

The Sum of the Squared Errors (SSE) Residual to Least Squares Fit

The least squares method of estimating the parameters of a probability density function (see Appendix A) aims at minimizing the sum of squared differences between the frequency of each class of a data histogram and the average of the hypothesized distribution for that class. After the fitting process is completed, one may expect the residual sum of squares to be small for a close fit. When data samples belonging to a particular PDF are fit to different PDF's, including the parent PDF, the value of the residual SSE might be expected to be the smallest, on the average, for the parent PDF.

Limitations on δ, D_o and SSE as Identifiers of PDF's

δ, D_o and SSE represent the errors between the sample and the fit. If the values of these statistics do not differ greatly when data are fit to different PDF's they may not serve as effective PDF identifiers. This will be verified by simulation experiments.

The Chi-Square Goodness-of-Fit Test

The test statistic, χ^2_o, of the chi-square goodness-of-fit test is given by,

$$\chi^2_o = \sum_{i=1}^{N} \frac{(O_i - n\bar{P}_i)^2}{n\bar{P}_i} \qquad (4.1)$$

in which O_i = Observed frequency in the ith class

$n\bar{P}_i$ = expected frequency of its class interval given by the product of sample size n and the probability, $\bar{P}_i$, for the ith class interval (see Figure A.1)

N = the number of class intervals of data histogram

It can be shown that the χ^2_o approximately follows chi-square distribution with N-γ-1 degrees of freedom when the sample is, in fact, from the hypothesized distribution. The number of parameters of the hypothesized distribution estimated by sample statistics is given by γ which is 2 in the present case. When $\delta = P(\chi^2 \geq \chi_o^2)$, the probability of exceeding the observed coefficient χ^2_o, falls below a certain value α (usually 0.05), the hypothesis that the random variable conforms to the hypothesized density is rejected. Based on this procedure, one may expect that δ is a measure of closeness of fit and that the larger the value of δ (the maximum value is of course unity) the closer the hypothesized density is expected to fit the data sample. In fact, χ^2_o (see Equation 4.1) will be zero for a perfect fit and the value of δ will

be unity. When data samples belonging to a particular PDF are fit to various PDF's, including the parent PDF, the values of δ may, on the average, be expected to be the largest for the parent fit. Thus, the largest value of δ will be examined as a possible criterion to discriminate PDF's.

The Kolmogorov-Smirnov (K-S) Goodness-of-Fit Test

The K-S goodness-of-fit test concentrates on the deviations between the hypothesized CDF, $F(x)$, and the observed or empirical CDF given by

$$F^*(X^{(i)}) = \frac{i}{n} \tag{4.2}$$

in which $X^{(i)}$ is the ith largest observed value in the random sample of size n. Consider the statistic

$$D_o = \max_{i=1}^{n} \left(\left| F^*(X^{(i)}) - F(X^{(i)}) \right| \right) = \max_{i=1}^{n} \left(\left| \frac{i}{n} - F(X^{(i)}) \right| \right) \tag{4.3}$$

D_o is the largest of the absolute values of the n differences between the hypothesized CDF and the empirical CDF evaluated at the observed values in the sample. Distribution of D_o is independent of the hypothesized distribution and the hypothesis that X has a specified distribution is accepted if

$$D_o \leq c \tag{4.4}$$

Criterion Based on Tolerance Limits

Statistical tolerance limits are boundaries between which a stated proportion of the population is expected to lie with respect to some measurable characteristic (Natrella, 1963). For a given data sample whose population is unknown, if tolerance limits are evaluated by fitting various PDF's, it is hypothesized that tolerance limits based on the parent PDF may show some distinct characteristics. For example, the 90% upper tolerance limit of an estimate of the 100-year hydrologic event based on the "best" PDF may be the highest (or the lowest) of all such values computed compared to other PDF's. Such characteristics may be evaluated from simulation experiments and the possibility of developing such a criterion to discriminate PDF's may be examined.

A Description of Numerical Experiments

It was for this purpose that a series of numerical simulation experiments were conducted. The five PDF-discriminating criteria presented above were applied to simulated data samples to evaluate their effectiveness in distinguishing the parent probability density function. The second objective of the numerical experiments was to search for trends in discrepancies between sample moments (particularly, the variance) and the moments of the density function fit by ML/LS/MCS methods when samples from some specific PDF are fit to other PDF's. The specific questions addressed were

1. Will the variance of a PDF fitted to a sample by a shape fitting method be, on the average, closer to the sample

variance when the fit is based on the parent PDF as opposed to some other PDF?

2. What is the nature of discrepancies between the variance of a PDF fitted by ML/LS/MCS and the sample variance when, a) LN samples are fit to GA and GU, b) GA samples are fit to LN and GU and c) GU samples are fit to LN and GA? How do such discrepancies affect predictions based on ML/LS/MCS?
3. Will the statistics of chi-square and Kolmogorov-Smirnov goodness-of-fit tests or the sum of squared errors (SSE) of a least squares fit serve as a device to identify the parent distribution if samples from a specific population are fit to various distributions?
4. How do the statistical tolerance limits at a certain confidence level (see Appendix A for a description of the method) vary if the data of a specific population are fit to various distributions? Will they, in any way, identify the parent distribution?

Selection of Density Functions for Numerical Experimentation

The lognormal (LN), gamma (GA) and the Gumbel (GU) (all two-parameter) distributions were selected for use in the numerical simulation.

Lognormal and Gumbel distributions are widely used by hydrologists in frequency analyses. The gamma distribution is also being increasingly used for studying the probability of occurrence of hydrologic events (Cruff and Rantz, 1965, Yevjevich, 1972). Though the three distributions may be made more flexible by including a so-called "shifting parameter",

two parameter distributions were retained in this study to avoid possible numerical difficulties which might arise in higher dimensional parameter models. Moreover, Markovic (1965) showed the three parameter gamma and lognormal distributions to have no significant advantage when dealing with hydrologic variates such as rainfall or runoff over the two parameter distributions.

The essential properties of the two parameter lognormal, the two parameter gamma and the two parameter Gumbel distributions (henceforth referred to simply as lognormal or LN, gamma or GA, and Gumbel or GU distributions) are summarized in Appendices B, C and D respectively.

Appendix E describes the schemes by which lognormal, gamma and Gumbel pseudorandom variates were generated.

In developing simulation runs, the population parameters of the distributions were always chosen to generate the equivalent dimensionless variates K_i, given by Equation 3.1. In other words, the population mean of the random variates generated was always unity. This was achieved simply by choosing the value of one of the parameters and computing the other by using Equations 3.6, 3.7, and 3.8 for lognormal, gamma and Gumbel distributions, respectively. To choose an appropriate parameter range for developing the simulation runs, stream flow data gathered from 67 gauging stations located throughout the United States (see Appendix H) were fit to the three distributions chosen for this study. The variance S_k^2 of the dimensionless data (K_i) at the 67 stations ranged from .058 to 2.494. However, 58 of the 67 stations have a variance less than 0.7 and there were only two stations with a variance larger than 1.0. The program will not evaluate least square parameters

bias was eliminated when the weighted LS was used (see the results of MCS method which is equivalent to LS method with a weight exponent of 1.0. See Appendix A for a description of LS method).

The means of Gumbel Samples at higher variance (σ_k^2 = 0.41 and 0.73) were positively biased (see column (2), Table 5.2) because negative variates were always discarded. Such a positive bias in the mean of the samples and the high dissimilarities in GU and LN/GA PDF's at σ_k^2 = 0.73 (see Figures 3.1 through 3.3) caused a very high positive bias in $\mu_{F,GU.LN/(ML/LS/MCS)}$ and $\mu_{F,GU.GA/(ML/LS/MCS)}$.

Discrepancies in the Second Moment

In Chapter III it was shown that the GA-PDF was always different from the LN while the GU-PDF resembled LN when σ_k^2 = 0.13 and GA when σ_k^2 = 0.32. The differences in the shapes of various PDF's can be used to explain the discrepancies between S_k^2 and σ_F^2 (most of these differences are discussed in Chapter III).

GA and GU PFD's Fit to LN Data

Figures 3.1 through 3.3 show that, in general, the main portion of the GA density curves are located closer to the origin than LN density curves at given σ_k^2 = 0.2 to 0.7. GU density curves in the range of σ_k^2 = 0.2 to 0.7 also are located closer to the origin than LN densities except for the fact that a portion of the rising limbs of GU densities lies in the negative range of K. Hence, with reference to the origin, the position of high variance LN densities is occupied by relatively low variance GA and GU densities. Since shape fitting methods fit the overall shape of the sample distribtuion, from the

above occurrences one might, in general, expect that if LN data at any σ_k^2 were fit to GA, or if LN data at σ_k^2 of about 0.2 or greater were fit to GU by ML/LS/MCS, the resulting fits would have a variance lower than the sample variance. This phenomenon is evident from Table 5.2. Table 5.2 shows that, with the exception of $\sigma^2_{F,LN.GU/(LS/MCS)}$ at $\bar{S}_k^2$ = 0.096, $\sigma^2_{F,LN.GU/(LS/MCS)}$ and $\sigma^2_{F,LN.GA/(ML/LS/MCS)}$ are always less than the variance of the data. At $\bar{S}_k^2$ = 0.63, which is the largest variance for LN data investigated, $\sigma^2_{F,LN.GA(ML/LS/MCS)}$ and $\sigma^2_{F,LN.GU/MCS}$ were found to be about 65% of the sample variance. The LS fit (ϕ = 0.00), in general, showed a larger discrepancy than ML/MCS for samples with $\bar{S}_k^2$ = 0.189 and above. However, such large discrepancies in LN fits will, if anything, serve to better discriminate the parent PDF; $\sigma^2_{F,LN.GA/LS}$ and $\sigma^2_{F,LN.GU/LS}$ are found to be less than 50% of the sample variance when $\bar{S}_k^2$ = 0.63.

At $\bar{S}_k^2$ = 0.096, $\sigma^2_{F,LN.GU/(LS/MCS)}$ did not show any appreciable discrepancy from the sample variance. This is because the LN and GU distributions do not differ much in shapes at $\sigma_k^2 \simeq 0.1$ (see Figures 3.1 and 3.3).

The foregoing discrepancies suggest the following conclusions:

a) When the data sample is from a LN PDF and the sample was fit to a GA PDF, the ML/LS/MCS predictions were, on the average, smaller than the predictions by the method of moments (see Table 5.3 for discrepancies in 100-year predictions). The values of $\bar{K}_{S100,LN.GA/ML}$ were found to be 96% and 86% of the MO predictions at S_k^2 = 0.173 and 0.630, respectively.

for a gamma fit if the distribution becomes exponential, i.e., when S_k^2 is larger than or equal to 1.0. For about 60 stations the LN parameter σ_y had a range of about 0.25 to 0.7, parameters of the gamma distribution had a range of about 2 to 11 and the dispersion parameter a of the Gumbel distribution had a range of about 1.5 to 4.0. Parameter values of these ranges were utilized in developing the simulation runs.

In an attempt to answer questions 1 through 3, three hundred random samples, each of size 100, were examined. These samples were drawn from LN, GA and GU populations (100 samples from each population) and covered a wide range of σ_k^2. All 300 samples were fit to LN, GA and GU PDF's by the methods of MO, ML (except GU fit), LS and MCS. To answer question 4, one sample of size 100 was selected from each of the three populations and the 90% upper tolerance limits of the 100-year events ($\overline{K}_{.99u,.90}$; see Appendix A) were evaluated. Chapter V represents and analyzes the results.

CHAPTER V

ANALYSIS OF RESULTS OF NUMERICAL EXPERIMENTS

This chapter describes the numerical simulation experiments outlined in Chapter IV. The procedure consisted of the following steps:

1. For a given set of parameter values (α,β)(the two parameters in a PDF will be designated as α and β, i.e., $\theta_1 = \alpha$, $\theta_2 = \beta$) generate on a digital computer 25 samples of size 100 from a LN distribution.
2. Fit each sample from Step 1 to a) a LN PDF, b) a GA PDF, and c) a GU PDF. Each type of PDF was fit by four methods (MO, LS, ML and MCS), and each fit yielded estimates (A, B) of the parameters (α,β). Each fit also yielded values of the mean and variance (μ_F, σ_F^2) of the fitted PDF's. (The exeption was that ML estimates of the GU parameters were not made because of limited computer time). Chi-square and K-S goodness-of-fit tests were made of the LS fit.
3. The values of K_i corresponding to $CDF(K_i) = 0.90, 0.96, 0.98, 0.99, 0.995, 0.998$, and 0.999 were determined for each PDF fitted in Step 2. (If the random variates are assumed to represent an annual series, these K_i would be associated with return periods of 10, 25, 50, 100, 200, 500 and 1000 years, respectively).

4. Compute the mean values ($\bar{A}$, $\bar{B}$) of 25 sample estimates in Step 2 for each fit by each method (i.e., ($\bar{A}$, $\bar{B}$) for LN PDF by ML, etc.)
5. Compute the mean and variance of each K_i determined in Step 3.
6. Repeat Steps 1 - 5 for 3 additional parameters (α,β).
7. Repeat Steps 1 - 6 using samples from GA distributions.
8. Repeat Steps 1 - 6 using samples from GU distributions.

The computer program which was originally formulated for study of LS method (see Appendix A) evaluates the following major steps:

a. For a given set of parameter values (α,β) generate a sample of specified size from LN, GA or GU distribution.
b. Arrange data into a histogram.
c. Fit the sample to the specified PDF (LN, GA or GU) by MO, ML (GU excepted) and weighted LS (using the given weight exponent ϕ. See Section II, Appendix A, for meaning of ϕ).
d. Perform chi-square and K-S test for normality on LS errors.
e. Perform chi-square and K-S goodness-of-fit tests on LS fit.
f. Determine the values of K_i corresponding to CDF (K_i) = 0.90, 0.96, 0.98, 0.99, 0.995, 0.998, and 0.999 for fit by LS and ML (MO in case of GU).
g. Repeat Steps a. through f. twenty times.
h. Summarize all results.

In this study a sample size of 100 (Step a) was chosen. ϕ = 0.0 in Step c is equivalent to LS method and ϕ = 1.0 is equivalent to MCS

method. Steps a) through h) are called a 'computer run.' However, the reader may note that because of limitations in Step c, if data samples were required to be fit to all three PDF's (LN, GA and GU) by MO, ML, LS and MCS, 6 'computer runs' would be necessary. These 6 runs will have the same data samples.

In all, 72 'computer runs' were made for the various analyses presented in this study. However, for convenience the sets of 6 'computer runs' which have the same data samples in each run (one run for each of three PDF's, each fit with $\phi = 0.0$ and $\phi = 1.0$) were designated by a single run number in the subsequent pages. (See also Appendix G for a description of computer runs. Appendix G lists all computer runs made for this study and the study on LS method. The total number of runs made for the two studies was 214. The findings of the study on LS method are summarized separately in Appendix A.)

Table 5.1 summarizes the values of the parameters used in the simulation runs and the population K values for various return periods for each parameter set chosen. Appendix G lists for each simulation run the parameter values, sample size, parent PDF, and the PDF to which the data were fit.

Study No. 1: Discrepancies in Moments of the Fitted Distribution and Predictions

The mean and variance of a fitted PDF depends on the choices of type of PDF and the parameter estimation method as well as on the characteristics of the sample. The value of the random variable

Table 5.1. Population Parameters used for Simulation Experiments

Run Series	Distribution and Population Parameters		σ_K^2	Population $Q/\bar{Q} = K$ for Return Periods (yrs) 10	25	50	100	200	500	1000
	Log Normal* $\alpha = \mu_y$	$\beta = \sigma_y$								
1LN	-.045	0.3	.0942	1.404	1.616	1.770	1.921	2.070	2.267	2.415
2LN	-.080	0.4	.1735	1.541	1.859	2.099	2.341	2.586	2.918	3.177
3LN	-.125	0.5	.2840	1.675	2.117	2.464	2.824	3.199	3.720	4.136
4LN	-.245	0.7	.6323	1.919	2.665	3.295	3.488	4.748	5.867	6.805
	Gamma* $\alpha = C$	$\beta = D$								
1GA	3	3	.3333[1]	1.774	2.200	2.505	2.802	3.091	3.465	3.743
2GA	5	5	.2000	1.599	1.902	2.116	2.321	2.519	2.772	2.959
3GA	7	7	.1429	1.505	1.749	1.919	2.082	2.237	2.435	2.580
4GA	10	10	.1000	1.421	1.615	1.751	1.878	2.000	2.154	2.266
	Gumbel* $\alpha = a$	$\beta = u$								
1GU	1.5	.6152	.7311	2.115	2.748	3.216	3.682	4.146	4.758	5.220
2GU	2.0	.7114	.4112	1.837	2.311	2.662	3.011	3.359	3.818	4.165
3GU	3.0	.8076	.1828	1.558	1.874	2.108	2.341	2.573	2.879	3.110
4GU	4.0	.8569	.1028	1.420	1.657	1.832	2.007	2.181	2.410	2.584

* See Table 3-1 for the equations of LN, GA, and GU distributions.

[1] The choice of parameter values was based on fitting of PDF's to real data for which (α, β) of GA varied from 2 to 11. For simulation, values of 3, 5, 7, and 10 were arbitrarily chosen for α and β of GA. This resulted in a variance range of 0.100 to 0.333.

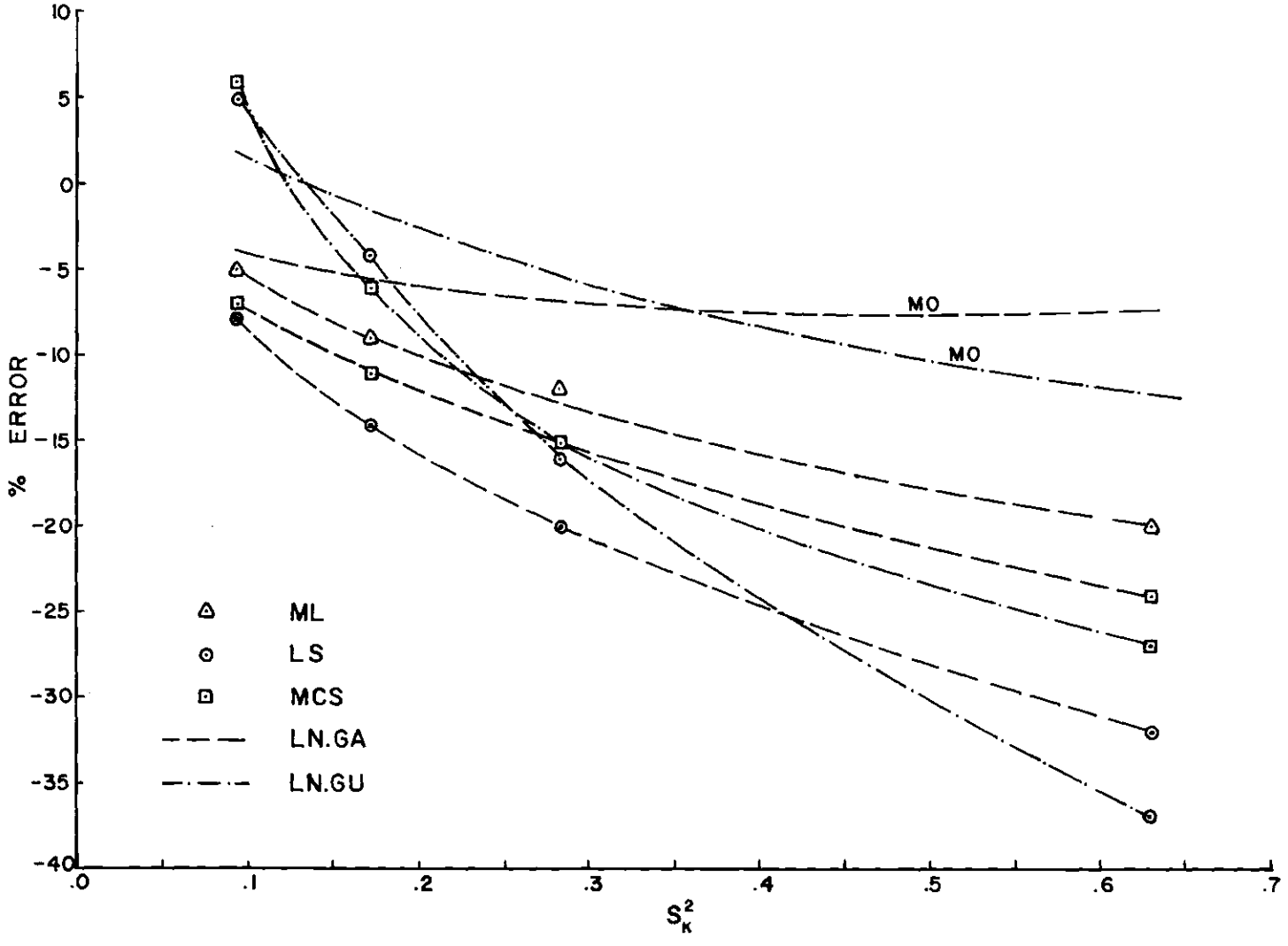


Figure 5.1 Errors in K_{S100} when GA and GU PDF's are fit to LN samples

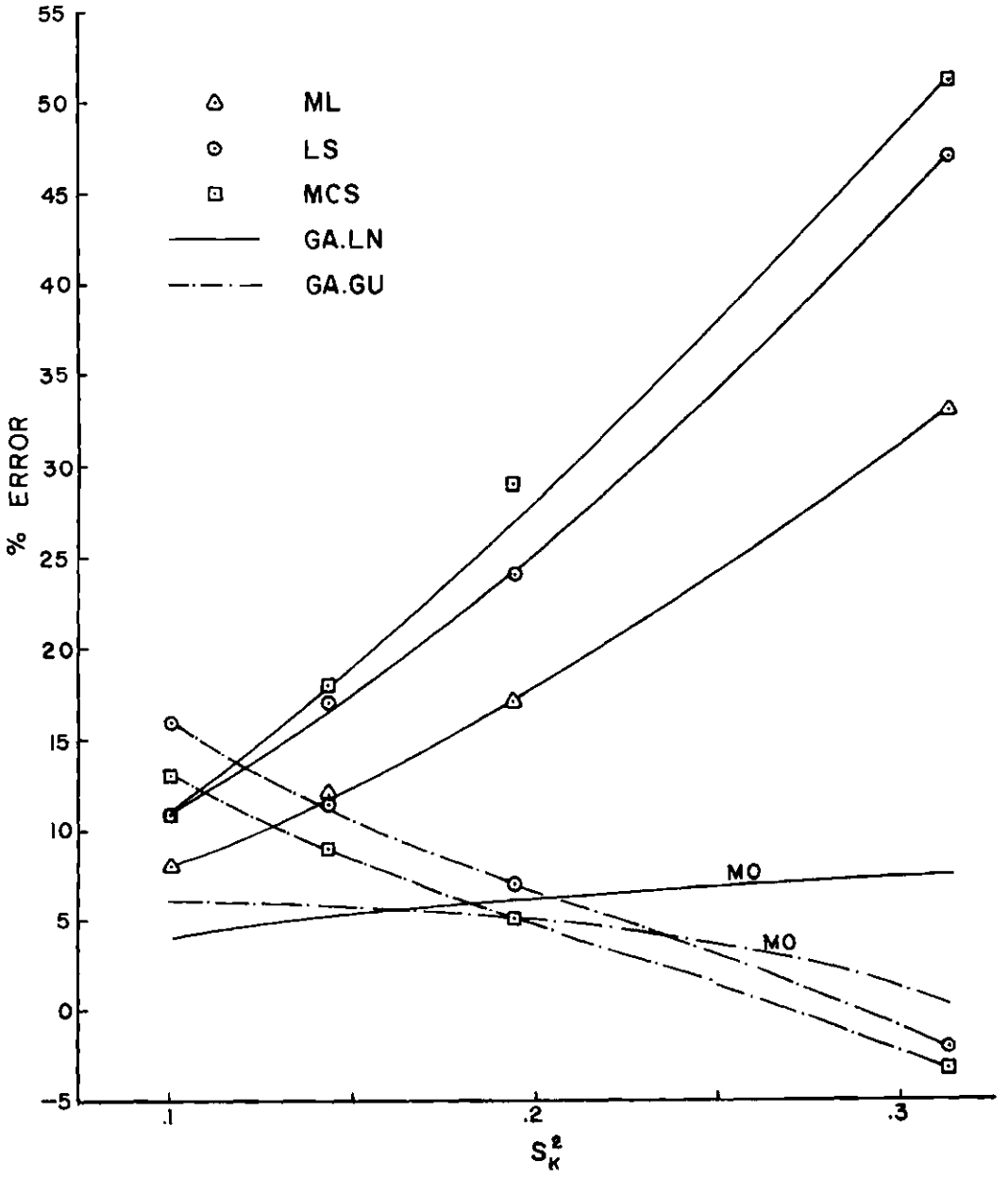


Figure 5.2 Errors in K_{S100} when LN and GU PDF's are fit to GA samples

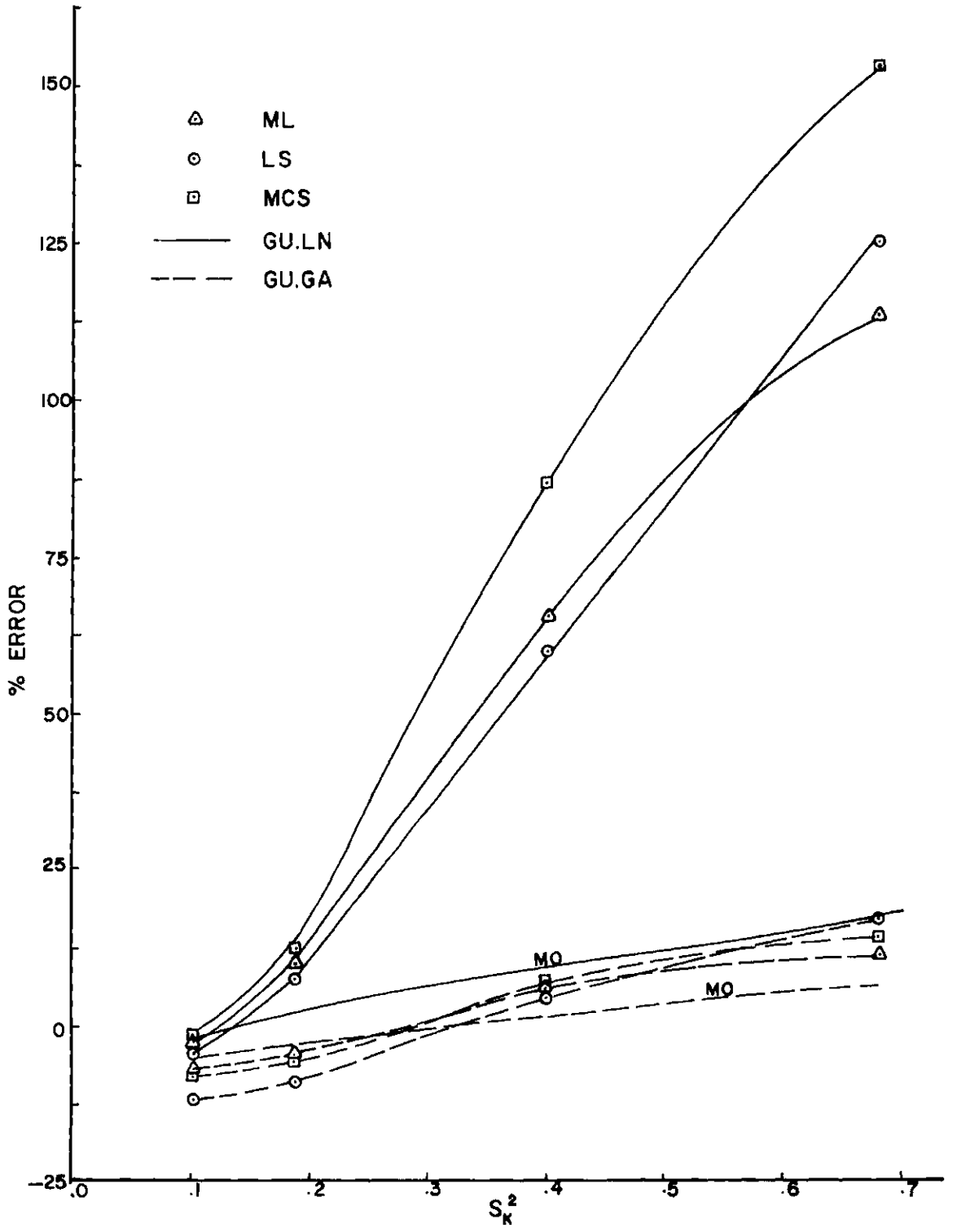


Figure 5.3 Errors in K_{S100} when LN and GA PDF's are fit to GU data

corresponding to given percentile of the CDF likewise depends on the same choices and characteristics. To study systematically the effects on mean and variance and on the K-values for various percentiles, data from a specific PDF are fit to various PDFs by ML, LS, MCS, and MO. Simulation runs were designed in which samples from a given population (LN, GA or GU) were fit to the three PDFs by MO, LS, ML and MCS estimation methods. (The term "predictions" is used hereafter to mean the estimated value of the random variable associated with a given probability of exceedence).

The average values of the parameters ($\bar{A}$, $\bar{B}$) were determined from each run (of 25 samples) and these values were used to determine μ_F and σ_F^2. The average value of the sample mean $\bar{M}_R$ and the average of the sample variance, $\bar{S}_K^2$, were also computed. These average values are compared in Table 5.2.

Table 5.2 shows that, in general, the moments of the PDF's fit by ML/LS/MCS are closer to the sample moments when data are fit to the parent distribution than when the data are fit to other distributions and that the differences are more pronounced as the variance of the data increases. The principal discrepancies may be summarized as follows:

Discrepancies in the First Moment

The first moment is not affected much by the fitting method or by the type of distribution (with the exception of Gumbel data at high variance, $\sigma_k^2 > 0.3$). This is particularly true with the ML method. However, $\mu_{F,LN.GA/LS}$ and $\mu_{F,LN.GU/LS}$ showed a negative bias, but the

Table 5.2. Comparison of the Sample Moments and Moments of the Fitted PDF When Data of a Given Population are Fit to Different PDF's

(Population Mean μ_K = 1.0; Sample Size = 100)

Data	$\bar{M}_K$	σ_K^2	$\bar{S}_K^2$	Fitted PDF	Moments of the Fitted PDF Based on Mean Parameters A&B						Run No.
					LS		ML		MCS		
					μ_F	σ_F^2	μ_F	σ_F^2	μ_F	σ_F^2	
LN	1.000	.094	.096	LN	1.003	.094	.999	.093	1.010	.101	1LN5
				GA	.997	.080	.996	.088	1.001	.089	
				GU	1.013	.102	*	*	1.013	.105	
	1.012	.173	.189	LN	1.002	.168	1.010	.181	1.021	.195	2LN5
				GA	.958	.127	1.008	.160	1.010	.164	
				GU	.984	.154	*	*	1.010	.169	
	1.016	.284	.315	LN	1.005	.275	1.014	.299	1.030	.331	3LN4
				GA	.937	.182	1.010	.247	1.013	.256	
				GU	.946	.205	*	*	1.003	.243	
	.984	.632	.630	LN	.973	.571	.978	.577	1.006	.646	4LN4
				GA	.868	.289	.974	.411	.980	.416	
				GU	.808	.265	*	*	.923	.413	
GA	.994	.100	.099	LN	1.030	.115	.996	.111	1.016	.128	4GA5
				GA	1.000	.095	.993	.099	1.001	.105	
				GU	1.038	.125	*	*	1.020	.131	
	.989	.143	.143	LN	1.037	.188	.992	.168	1.031	.208	3GA6
				GA	.989	.141	.989	.142	1.000	.153	
				GU	1.019	.173	*	*	1.015	.175	
	.984	.200	.194	LN	1.052	.287	.991	.254	1.056	.344	2GA5
				GA	.986	.194	.986	.195	1.004	.218	
				GU	.998	.222	*	*	1.010	.231	
	.978	.333	.326	LN	1.100	.636	1.000	.518	1.101	.727	1GA5
				GA	.970	.318	.973	.321	.996	.351	
				GU	.935	.318	*	*	.980	.343	
GU	1.000	.103	.104	LN	.985	.093	.999	.099	1.007	.105	4GU5
				GA	.966	.080	.997	.094	1.000	.095	
				GU	.994	.098	*	*	1.007	.105	
	1.014	.183	.191	LN	1.041	.217	1.018	.222	1.046	.253	3GU5
				GA	.993	.162	1.012	.179	1.022	.189	
				GU	1.012	.188	*	*	1.025	.203	
	1.031	.411	.398	LN	1.221	.833	1.100	.952	1.244	1.324	2GU4
				GA	1.050	.389	1.023	.426	1.064	.481	
				GU	1.006	.388	*	*	1.031	.449	
	1.146	.731	.682	LN	1.529	2.501	1.302	2.549	1.554	3.884	1GU5
				GA	1.224	.799	1.143	.760	1.200	.864	
				GU	1.038	.658	*	*	1.072	.772	

* Not evaluated

bias was eliminated when the weighted LS was used (See the results of MCS method which is equivalent to LS method with a weight exponent of 1.0. See Appendix A for a description of LS method).

The means of Gumbel samples at higher variance (σ_k^2 = 0.41 and 0.73) were positively biased (see column (2), Table 5.2) because negative variates were always discarded. Such a positive bias in the mean of the samples and the high dissimilarities in GU and LN/GA PDF's at σ_k^2 = 0.73 (See Figures 3.1 through 3.3) caused a very high positive bias in $\mu_{F,GU.LN/(ML/LS/MCS)}$ and $\mu_{F,GU.GA/(ML/LS/MCS)}$.

Discrepancies in the Second Moment

In Chapter III it was shown that the GA-PDF was always different from the LN while the GU-PDF resembled LN when $\sigma_k^2 \simeq 0.13$ and GA when $\sigma_k^2 \simeq 0.32$. The differences in the shapes of various PDF's can be used to explain the discrepancies between S_k^2 and σ_F^2 (most of these differences are discussed in Chapter III).

GA and GU PDF's fit to LN Data

Figures 3.1 through 3.3 show that, in general, the main portion of the GA density curves are located closer to the origin than LN density curves at given $\dot{X}$. σ_k^2 density curves in the range of 0.2 to 0.7 also are located closer to the origin than LN densities except for the fact that a portion of the rising limbs of GU densities lies in the negative range of K. Hence, with reference to the origin, the position of high variance LN densities is occupied by relatively low variance GA and GU densities. Since shape fitting methods fit the overall shape of the sample distribution, from the above occurrences one might, in general, expect that if LN

data at any σ_k^2 were fit to GA, of if LN data at σ_k^2 of about 0.2 or greater were fit to GU by ML/LS/MCS, the resulting fits would have a variance lower than the sample variance. This phenomenon is evident from Table 5.2. Table 5.2 shows that, with the exception of $\sigma^2_{F,\ LN.GU/(LS/MCS)}$ at $\bar{S}_k^2 = 0.096$, $\sigma^2_{F,\ LN.GU/(LS/MCS)}$ and $\sigma^2_{F,\ LN.GA/(ML/LS/MCS)}$ are always less than the variance of the data. At $\bar{S}_k^2 = 0.63$, which is the largest variance for LN data investigated, $\sigma^2_{F,\ LN.GA/(ML/LS/MCS)}$ and $\sigma^2_{F,\ LN.GU/MCS}$ were found to be about 65% of the sample variance. The LS fit ($\phi = 0.00$), in general, showed a larger discrepancy than ML/MCS for samples with $\bar{S}_k^2 = 0.189$ and above. However, such large discrepancies in LN fits will, if anything serve to better discriminate the parent PDF; $\sigma^2_{F,\ LN.GA/LS}$ and $\sigma^2_{F,\ LN.GU/LS}$ are found to be less than 50% of the sample variance when $\bar{S}_k^2 = 0.63$.

At $\bar{S}_k^2 = 0.096$, $\sigma^2_{F,\ LN.GU/(LS/MCS)}$ did not show any appreciable discrepancy from the sample variance. This is because the LN and GU distributions do not differ much in shapes at $\sigma_k^2 \approx 0.1$ (see Figures 3.1 and 3.3).

The foregoing discrepancies suggest the following conclusions:

a) When the data sample is from a LN PDF and the sample was fit to a GA PDF, the ML/LS/MCS predictions were, on the average, smaller than the predictions by the method of moments (see Table 5.3 for discrepancies in 100-year predictions). The values of $\bar{K}_{S100,\ LN.GA/ML}$ were found to be 96% and 86% of the MO predictions at $S_k^2 = 0.173$ and 0.630, respectively.

b) When the data sample was from a LN PDF and the sample was fit to a GU PDF, at low variances, say S_k^2 up to about 0.15, the ML/LS/MCS predictions were approximately same as MO predictions. At higher variances the ML/LS/MCS predictions were smaller than the MO predictions. The values of $K_{S100,LN.GU/MCS}$ were found to be 98% and 88% of the MO predictions at sample variances of 0.173 and 0.630, respectively. The LS predictions were much lower.

LN and GU PDF's Fit to GA Data

In the foregoing section it was mentioned that the position of high variance LN densities, with reference to origin, was occupied by relatively low variance GA densities. Under such circumstances, if LN is fit to GA data by shape fitting methods, the LN fit, in order to approximate the GA distribution, may be expected to have a larger variance than the GA data. The results presented in Table 5.2 corroborate the above. $\sigma^2_{F,GA.LN/(ML/LS/MCS)}$ was, on the average, found to be larger than S_k^2 in the range of S_k^2 investigated ($\bar{S}_k^2$ = 0.099 to 0.326), and $\sigma^2_{F,GA.LN/LS}$ was found to be even larger still. The above trend in the variance of LN fit to GA data is expected to persist for higher values of $\bar{S}_k^2$ because the noted differences in the shapes of these distributions increases with increase in variance (see Figures 3.1 and 3.2).

The GU distribution is similar to LN distribution at low variances ($\sigma_k^2 < 0.2$ see Figures 3.1 through 3.5) and then attains a close resemblance with GA in the variance range of 0.3 to 0.4 (see

Figures 3.2, 3.3 and 3.6). When $\sigma_k^2 = 0.5$ or above GU becomes flatter with tails thicker compared to LN or GA densities. Due to the above mentioned differences in the shapes of the GA and GU density curves, $\sigma^2_{F,GA.GU/(LS/MCS)}$ is expected to be larger when $S_k^2 < 0.3$, approximately the same when $S_k^2 = 0.3$ to 0.4 and smaller when $S_k^2 > 0.4$. (The last mentioned occurrence may be expected because the dissimilarities between the GA and GU densities are similar to the dissimilarities between LN and GA (or GU) when $\sigma_k^2 > 0.4$). The results shown by Table 5.2 confirm the above expectations. Table 5.2 shows that the variance of the fitted PDF, $\sigma^2_{F,GA.GU/LS}$, is 26% and 14% larger than the sample variance at $\bar{S}_k^2 = 0.099$ and 0.194, respectively, but they become equal at $\bar{S}_k^2 = 0.326$. (The variances of MCS method were slightly larger than by LS because MCS gives more weight to the tail of the distribution as explained in Appendix A).

The foregoing discrepancies between the variance of the fitted PDF and the sample variance suggest the following conclusions:

a) When the data sample is from a GA distribution and the sample is fit to a LN distribution the ML/LS/MCS predictions of future occurrences will, on the average, be larger than the MO predictions (see Table 5.3 for discrepancies in 100-year predictions). The $\bar{K}_{S100,GA.LN/ML}$ were found to be 4% and 25% larger than the LN-moment predictions at $\bar{S}_k^2 = 0.099$ and 0.326, respectively. The $\bar{K}_{S100,GA.LN/(LS/MCS)}$ were found to be much larger.

b) When the sample is from a GA distribution and the sample is

Table 5.3. Discrepancies in LS, ML, MC and MO Predictions When Data of Given Population Are Fit to Different PDF's

Sample Size = 100

Data	$\overline{S^2_K}$	Fitted PDF	K_{100} ($\overline{S^2_K}=\sigma^2_K$) (Fig 3.9f)	$\overline{K}_{S100}$ MO	LS	ML	MCS	RATIO $\overline{K}_{S100}$ LS/MO	ML/MO	MCS/MO	Run No.
LN	.094	LN	1.92	1.92	1.93	1.92	1.97	1.005	1.000	1.026	1LN5
		GA	1.85	1.84	1.77	1.83	1.84	.962	.995	1.000	
		GU	1.97	1.96	2.03	*	2.04	1.036	*	1.041	
	.173	LN	2.34	2.40	2.33	2.39	2.47	.971	.996	1.029	2LN5
		GA	2.20	2.27	2.01	2.18	2.20	.885	.960	.969	
		GU	2.30	2.36	2.23	*	2.31	.945	*	.979	
	.284	LN	2.82	2.93	2.82	2.90	3.04	.962	.990	1.038	3LN4
		GA	2.62	2.73	2.25	2.54	2.58	.824	.930	.945	
		GU	2.66	2.75	2.38	*	2.57	.865	*	.935	
	.630	LN	3.99	3.86	3.88	3.86	4.07	1.005	1.000	1.054	4LN4
		GA	3.69	3.59	2.62	3.10	3.10	.730	.864	.864	
		GU	3.49	3.40	2.45	*	2.98	.721	*	.876	
GA	.099	LN	1.96	1.94	2.07	2.02	2.12	1.067	1.041	1.09	4GA5
		GA	1.88	1.86	1.87	1.87	1.91	1.005	1.005	1.027	
		GU	1.99	1.97	2.16	*	2.16	1.096	*	1.096	
	.143	LN	2.19	2.17	2.45	2.32	2.54	1.129	1.069	1.171	3GA6
		GA	2.08	2.07	2.09	2.08	2.15	1.010	1.005	1.039	
		GU	2.19	2.17	2.34	*	2.34	1.078	*	1.078	
	.194	LN	2.46	2.41	2.89	2.70	3.11	1.199	1.120	1.290	2GA5
		GA	2.32	2.27	2.33	2.30	2.41	1.026	1.103	1.062	
		GU	2.40	2.35	2.50	*	2.53	1.064	*	1.077	
	.326	LN	3.01	2.96	4.11	3.69	4.42	1.389	1.247	1.493	1GA5
		GA	2.80	2.75	2.80	2.78	2.93	1.018	1.011	1.065	
		GU	2.80	2.75	2.74	*	2.85	.996	*	1.036	
GU	.104	LN	1.98	1.97	1.91	1.95	1.99	.970	.990	1.010	4GU5
		GA	1.89	1.89	1.76	1.86	1.86	.931	.984	.984	
		GU	2.00	2.00	1.99	*	2.03	.995	*	1.015	
	.191	LN	2.38	2.42	2.56	2.59	2.74	1.058	1.070	1.132	3GU5
		GA	2.24	2.28	2.17	2.26	2.30	.952	.991	1.009	
		GU	2.34	2.37	2.39	*	2.44	1.008	*	1.030	
	.398	LN	3.29	3.24	4.80	4.94	5.94	1.481	1.525	1.833	2GU4
		GA	3.05	3.00	3.15	3.17	3.39	1.050	1.057	1.130	
		GU	3.01	2.99	3.02	*	3.18	1.010	*	1.064	
	.682	LN	4.25	4.15	8.10	7.89	9.70	1.952	1.901	2.337	1GU5
		GA	3.94	3.84	4.22	4.11	4.37	1.100	1.070	1.138	
		GU	3.68	3.71	3.60	*	3.84	.970	*	1.035	

* Not evaluated

fit to a GU distribution the LS, (ML where available) and the MCS predictions will be, compared to the predictions by the method of moments, larger at low variances ($S_k^2 < 0.3$), approximately the same in the variance range of 0.3 to 0.4, and lower when $S_k^2 > 0.4$. Table 5.3 shows the $\bar{K}_{S100}$ values based on MO, LS and MCS methods for GU fit of GA data in the variance range of 0.099 to 0.326. No simulation experiments were carried out with GA data for $\bar{S}_k^2 > 0.326$. Table 5.3 shows that $K_{S100,GA.GU/LS}$ is about 10% larger than $K_{S100,GA.GU/MO}$ at $\bar{S}_k^2 = 0.099$ and the MO and LS $\bar{K}_{S100}$ values become approximately equal at $\bar{S}_k^2 = .326$.

LN and GA PDF's Fit to GU Data

Under the foregoing two sub-headings the similarities and differences between the shapes of GU and LN and between GU and GA are discussed with references to a wide range of σ_k^2. The effect of these differences on the σ_F^2 of the PDF's fitted by shape fitting methods are also discussed. Without going into details, the discrepancies between the σ_F^2 based on ML/LS/MCS and the sample variance when GU data are fit to LN and GA distributions may be summarized as follows:

a) When the data sample is from a GU distribution and the sample is fit to a LN distribution, the σ_F^2 of ML/LS/MCS methods would be approximately the same as the sample variance up to a S_k^2 of about 0.15. At larger values of S_k^2, $\sigma^2_{F,GU.LN/(ML/LS/MCS)}$ would be, on the average, larger than S_k^2. The above occurrences are illustrated in Table 5.2.

At $\sigma_k^2 = 0.73$ ($\bar{S}_k^2 = .68$), where the GU and LN distributions are highly different, (see Figures 3.1 and 3.3) the $\sigma^2_{F,GU.LN/(ML/LS/MCS)}$ are found to be as high as (3.7/3.7/5.7) times the sample variance (see Table 5.2). The discrepancies in predictions are given by Table 5.3. Table 5.3 shows that $\bar{K}_{S100,GU.LN(ML/LS/MCS)}$ are higher by about 6% to 13% at $\bar{S}_k^2 =$ 0.19 and 90% to 134% at $\bar{S}_k^2 = 0.68$ compared to the $\bar{K}_{100,GU.LN/MO}$. The higher figures are from the MCS method.

b) When the data sample is from a GU distribution and the sample is fit to a GA distribution, $\sigma^2_{F,GU.GA/(ML/LS/MCS)}$ would be less than S_k^2 when S_k^2 is less than 0.3, approximately equal to S_k^2 when S_k^2 is about 0.3 to 0.4 and larger than S_k^2 when S_k^2 is greater than 0.4 (see the foregoing sub-head for an explanation of such an occurrence). However, Table 5.2 shows that the discrepancies that occur in σ_F^2 when modified GU data are fit to GA are not as large as those when GU data are fit to LN, particularly at larger variance ($\sigma_k^2 = 0.73$, $\bar{S}_k^2 = 0.68$). The above phenomenon indicates that the Gumbel data (with the variates in the negative range of k discarded) readily fit gamma distribution in a wide range of variance (0.2 to 0.7). The analysis of 100-year predictions based on GA fit to Gumbel data by different methods (Table 5.3) shows that the LS/ML/MCS predictions differ from the moment predictions only by a small fraction, which further illustrates the easy adaptability of modified GU data to

GA distribution in the range of variance studied ($\bar{S}_k^2$ = 0.10 to 0.68).

Based on this analysis, one might state that the modified Gumbel distribution would easily fit a lognormal PDF up to a variance of about 0.2, and then would fit a gamma distribution until it is no longer suitable for application to real world hydrologic data, i.e., when σ_k^2 is about 0.6 at which the negative tail of GU distribution reaches 5% of the total area. Thus, when hydrologic data are subjected to frequency analysis, it may not be necessary to consider the Gumbel distribution as the possible parent distribution because either the LN or the GA distribution will provide as good a fit as would the GU (in its range of applicability).

Results from Individual Samples of a Run

In Table 5.2 the values of variance of PDF's fitted by different methods were computed using the mean parameter values based on twenty five individual samples. In order to verify whether the results based on individual samples of a simulation run would also lead to the general conclusions discussed in the foregoing paragraphs, the variance ratio (σ_F^2/S_k^2) was calculated for each sample for some runs of Table 5.2 and in Table 5.4 are presented the sample mean and variance of (σ_F^2/S_k^2) for these runs. In Table 5.4 are also presented the ratios ($\sigma_F^2/\bar{S}_k^2$), in which σ_F^2 was calculated from ($\bar{A}$, $\bar{B}$), for each run. Table 5.4, like Table 5.2, also shows that the variance of PDF's fit by shape fitting methods would be, on the average, closer to the sample variance when data are fit to the parent PDF and it would differ otherwise depending on the

Table 5.4. Discrepancies in σ_F^2 Based on Individual Samples

(Number of Samples per Run = 25)

Data	σ_K^2	Fitted PDF	$\sigma^2/\bar{S}_K^2$ based on $(\bar{A}, \bar{B})$ ML	LS	Mean (σ_F^2/S_K^2) ML	LS	Var (σ_F^2/S_K^2) ML	LS	Run No.
LN	0.284	LN	.95	.87	.99	.94	.022	.090	3LN4
		GA	.78	.58	.82	.64	.011	.027	
		GU	*	.65	*	.70	*	.028	
	0.632	LN	.92	.91	1.04	*	.062	*	4LN4
		GA	.65	.46	.77	*	.024	*	
GA	0.200	LN	1.31	1.48	1.34	*	.051	*	2GA5
		GA	1.01	1.00	1.05	*	.010	*	
	0.333	LN	1.51	1.96	1.66	2.10	.165	.734	1GA5
		GA	.99	.98	1.05	1.05	.010	.059	
		GU	*	.98	*	1.03	*	.046	
GU	0.411	LN	*	2.09	*	2.45	*	2.899	2GU4
		GA	*	.98	*	1.11	*	.140	
		GU	*	.98	*	1.06	*	.091	

* Not evaluated

Table 5.5. Comparison of Ratios, σ_F^2/S_K^2 - GA Data

Sample Size = 100

Sample No	Run No 2GA5, σ_K^2 = 0.200			Run No 1GA5, σ_K^2 = 0.333				
		σ_F^2/S_K^2			σ_F^2/S_K^2			
	S_K^2	GA fit by ML	LN fit by ML	S_K^2	GA fit by LS	LN fit by LS	GA fit by ML	LN fit by ML
1	.154	.962	1.071	.331	1.191	2.573	1.119	1.894
2	.212	.970	1.063	.239	1.087	1.594	.997	1.325
3	.186	1.059	1.433	.290	.859	1.556	1.049	1.307
4	.217	1.154	1.492	.249	.847	1.282	.912	1.073
5	.170	1.124	1.404	.410	1.275	3.149	.964	1.492
6	.270	.938	1.192	.277	1.271	2.399	1.172	1.929
7	.170	1.029	1.183	.381	.683	1.527	.943	1.165
8	.192	1.024	1.213	.292	.895	1.540	1.132	1.856
9	.186	1.104	1.569	.318	.850	1.510	1.045	1.636
10	.227	.968	1.296	.334	.922	1.712	1.001	1.435
11	.157	1.222	1.642	.439	1.090	2.546	1.020	2.188
12	.186	1.088	1.439	.265	.770	1.237	1.012	1.250
13	.193	.910	1.141	.347	.995	1.751	1.015	1.646
14	.197	1.237	1.685	.278	1.180	2.552	1.092	1.471
15	.176	1.110	1.456	.448	1.355	3.569	1.052	2.226
16	.228	.962	1.197	.303	1.060	1.762	.990	1.477
17	.197	1.243	1.976	.425	.865	1.405	.890	1.472
18	.226	.989	1.155	.275	.914	1.636	1.041	1.474
19	.124	1.142	1.465	.248	1.117	1.847	1.220	1.766
20	.194	.991	1.176	.292	1.490	3.394	1.231	2.353
21	.198	1.068	1.407	.395	.905	1.373	.921	1.514
22	.188	1.050	1.327	.361	.931	1.777	1.014	1.638
23	.260	.842	1.041	.356	1.604	4.528	1.268	2.825
24	.171	1.046	1.240	.336	.754	1.304	.981	1.313
25	.175	1.009	1.161	.264	1.407	2.933	1.160	1.657

Table 5.6. Comparison of Ratios, σ_F^2/S_K^2 - LN Data

Sample Size = 100

Sample No	Run No 3 LN 4, σ_κ^2 = 0.284					Run No 4LN4, σ_κ^2 = 0.632		
		σ_F^2/S_κ^2					σ_F^2/S_κ^2	
	S_κ^2	LN fit by LS	GA fit by LS	LN fit by ML	GA fit by ML	S_κ^2	LN fit by ML	GA fit by ML
1	.350	.848	.602	.948	.758	.625	1.107	.752
2	.143	.853	.654	.992	1.035	1.117	.869	.566
3	.262	1.591	.923	1.189	.953	.628	1.033	.718
4	.313	.780	.555	.887	.792	.421	1.195	.942
5	.325	.793	.611	1.086	.819	.365	1.310	.967
6	.365	.708	.500	.930	.765	.468	.966	.811
7	.215	.810	.617	.969	.874	.647	.919	.679
8	.492	.906	.556	.877	.701	.591	.881	.732
9	.249	.991	.721	.996	.830	1.809	.503	.375
10	.285	.836	.589	.879	.800	.395	1.283	.891
11	.541	.517	.361	.666	.586	.985	.638	.528
12	.348	.978	.642	1.269	.892	.358	1.159	.919
13	.351	.964	.632	1.054	.828	1.196	.610	.475
14	.295	1.093	.730	1.038	.869	.720	1.274	.778
15	.199	1.476	.990	1.201	1.008	.470	1.253	.884
16	.332	.688	.526	.911	.742	.415	1.198	.877
17	.300	.718	.521	.966	.829	.502	1.325	.866
18	.368	.708	.492	.894	.750	.401	1.137	.843
19	.323	1.184	.763	1.137	.857	.471	.766	.748
20	.332	1.093	.684	1.001	.830	.583	.878	.699
21	.317	1.091	.729	.995	.804	.738	1.374	.775
22	.380	.406	.309	.647	.627	.481	1.035	.855
23	.229	.656	.520	.931	.829	.425	1.421	.966
24	.233	1.409	.910	1.178	.962	.603	.867	.693
25	.338	1.433	.837	1.093	.854	.338	1.088	.941

hypothesized PDF. (Particularly, when GA or GU data were fit to LN by ML or LS the mean (σ_F^2/S_k^2) was found to be even larger than $\sigma_F^2/\bar{S}_k^2$ for each run.

Tables 5.5 and 5.6 present a comparative statement of ratios $\sigma_F{}^2/S_k{}^2$ for GA and LN samples, respectively, for some runs of Table 5.4 (GU was omitted for simplicity). In Tables 5.5 and 5.6 the values of sample variance, S_k^2, are also shown for each sample of a run.

Table 5.5 shows that the ratio $\sigma_F{}^2/S_k{}^2$ is closer to unity when GA samples are fit to GA than to LN by ML/LS. Only one sample from each run (sample Nos. 4 and 23 of Runs 1GA5 and 2GA5, respectively) proved an exception. For these two samples the sample variance was found to be different (the difference is 24% to 30% of $\sigma_k{}^2$) from the population variance.

Table 5.6 shows that for 21 samples of Run 3LN4 and 15 samples of Run 4LN4 the ratio σ_F^2/S_k^2 is closer to unity when LN samples are fit to LN than to GA by ML/LS. Run 4LN4 had a relatively high population variance ($\sigma_k^2 = 0.632$) and the 10 samples for which $\sigma^2_{F,LN.LN/(ML/LS)}$ was not close to S_k^2 had their S_k^2 less than σ_k^2 by 21% to 44%. This shows that the statistical characteristics of the samples with high population variance are to be further investigated. However, for most natural data S_k^2 is less than 0.5 (see Chapter VI) and for samples within this range of S_k^2, in general, the ratio σ_F^2/S_k^2 was found to be closer to unity when data were fit to parent PDF by ML/LS.

Errors in Estimation of K_{S100} When the Correct Parent PDF Is Not Chosen

Table 5.3, a summary of 100-year predictions by different estimation methods from given data, illustrates a very important phenomenon for hydrologists who must estimate extreme events without knowing the correct distribution. The errors in not choosing the correct distribution are shown to be more severe when the computations are made by a so-called statistically superior method like ML than when the computations are made by the method of moments. As two of the many examples found on the Table, when the parent distribution is LN with variance 0.63 ($\tilde{S}_k^2$), the 100-year predictions from a fitted GA distribution were 3.59 and 3.10 for MO and ML fits, respectively, compared to the correct (based on LN population) value of 3.99. Similarly, 100-year predictions from a LN fit of a GA sample (variance = 0.33) were 2.96 and 3.69 for MO and ML fits, respectively, compared to the correct value (based on GA population) of about 2.80. In both cases, the error introduced by not choosing the correct distribution were larger when the ML method was used than when the MO method was used. The errors for situations in which the wrong distribution was chosen and at different variance levels are summarized in Table 5.7 for 100-year predictions. Errors in Table 5.7 are computed by using the formula,

$$\% \text{ Error} = \frac{(\text{Computed } K_{S100} - \text{population } K_{100})}{\text{population } K_{100}} \times 100 \qquad (5.1)$$

The errors introduced by choosing the wrong distribution are

Table 5.7. Errors Introduced in 100-Year Predictions By Different Estimating Methods When the Correct Distribution Is Not Chosen

Data	$\overline{S}_K^2$	Fit	% Error				Run No.
			MO	LS	ML	MCS	
LN	.094	GA	- 4	- 8	- 5	- 7	1LN5
		GU	+ 2	+ 5	**	+ 4	
	.173	GA	- 5	-14	- 9	-11	2LN5
		GU	- 2	- 4	**	- 6	
	.284	GA	- 7	-20	-12	-15	3LN4
		GU	- 6	-16	**	-15	
	.630	GA	- 7	-32	-20	-24	4LN4
		GU	-12	-37	**	-27	
GA	.099	LN	+ 4	+11	+ 8	+11	4GA5
		GU	+ 6	+16	**	+13	
	.143	LN	+ 5	+17	+12	+18	3GA6
		GU	+ 5	+12	**	+ 9	
	.194	LN	+ 6	+24	+17	+29	2GA5
		GU	+ 4	+ 7	**	+ 5	
	.326	LN	+ 8	+47	+33	+51	1GA5
		GU	0	- 2	**	- 3	
GU	.104	LN	- 2	- 4	- 3	- 2	4GU5
		GA	- 6	-12	- 7	- 8	
	.191	LN	+ 2	+ 7	+ 9	+12	3GU5
		GA	- 4	- 9	- 5	- 6	
	.398	LN	+ 8	+59	+65	+87	2GU5
		GA	0	+ 4	+ 6	+ 7	
	.682	LN	+12	+125	+113	+153	1GU5
		GA	+ 4	+17	+11	+14	

** Not computed

always smaller when MO method is used. Particularly to be noticed are the large errors introduced when LN is fit to high variance GA or GU data. The results presented in Table 5.7 are graphically shown by Figures 5.1, 5.2, and 5.3.

The 'Best Fit' Criterion

The findings of Study No. 1 indicate that close matching of sample moments to the moments of a PDF fitted by a shape fitting statistical estimation method may identify the probability density function of the best fit. Indeed, comparison of a single statistic, namely the variance of the sample to the variance of the PDF fitted by a shape fitting method, might suffice as an effective criterion. Mathematically, the 'best fit' criterion may be stated as follows:

A criterion for selecting an appropriate PDF, f(x), for fitting a random sample (X_i) of size n is to select that f(x) which makes the statistic

$$\frac{\text{Var } (X_i | f(x;\underset{\sim}{\theta}))}{\text{Var } (X_i)} \tag{5.2}$$

closest to unity, where

$$\text{Var } (X_i) = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1) \tag{5.3}$$

$$\text{Var } (X_i | f(x;\underset{\sim}{\theta})) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x;\underset{\sim}{\theta})dx \tag{5.4}$$

and the parameters of f(x), $\underset{\sim}{\theta}$, are estimated by a shape fitting method such as ML/LS/MCS and the mean value, μ, is estimated from

$$\mu = \int_{\infty} x f(x;\underset{\sim}{\theta})dx \tag{5.5}$$

$\overline{X}$ in equation 5.2 is given by

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{5.6}$$

ML method may be preferred to LS/MCS to compute Var. $(X_i | f(x;\theta))$ because unlike in LS/MCS the data do not have to be grouped in ML.

Study No. 2: PDF Discriminating Criteria Based on Statistics of Chi-Square and K-S Goodness-of-Fit Tests and LS Fit

For the PDF which best fits a sample, one would expect the statistics (i) $\delta=P(\chi^2>\chi_o^2)$ of chi-square test to be a maximum, (ii) D_o, the test statistic of Kolmogorov-Smirnov test to be a minimum, and (iii) SSE, the sum of squared errors residual to LS fit to be a minimum. To test whether the statistics δ, D_o or SSE are useful in identifying the proper distribution, numerical values of these three statistics were evaluated for the LS fit of each sample of the simulation runs mentioned in Table 5.2. In the evaluation of δ, the sample histogram organized for LS estimation (see Appendix A) was generally retained for chi-square tests also, but the observations in the tail of the distributions were lumped so that the expected frequency was at least 3 for the tail classes.

The simulation runs mentioned in Table 5.2 fit samples of a given PDF to the LN, GA, and GU PDF's. For each case, the values of δ, D_o and SSE obtained from LS fit were compared. Then the three criteria mentioned above were applied to decide which distribution best fit each sample. For example, assume that a sample of lognormal data is fit to the three distributions. Suppose it is observed that the value of δ is

the largest for LN fit and the values of D_o and SSE are minimum for GA and LN fits, respectively. Then it may be said that the data sample is closely (or best) fit by a LN distribution according to the chi-square test and Min SSE criterion, and that the GA distribution is indicated by the K-S test.

Tables 5.8 through 5.10 summarize the results for LN, GA and GU distributions, respectively. These tables show the number of cases indicated as the close (or best) fits by each discriminating criterion for each PDF when data samples of a particular density were fit to all three densities.

To test the validity of the above mentioned discriminating criteria, first the PDF's were compared by pairs then all three PDF's were simultaneously compared.

Tables 5.8 through 5.10 show that, in most cases, the three statistics, δ, D_o and SSE (residual to LS fit), identify the parent PDF when the sample is fit to various PDF's including the parent. Particularly, the results are better when fits to two PDF's are compared at a time (see parts a and b of Tables 5.8 through 5.10). However, in the σ_k^2 ranges in which LN and GU and GA and GU are close to each other (see Chapter III) identification of parent PDF from samples is not satisfactory (see Runs 2LN5 - Table 5.8b, and 4GU5 - Table 5.10a), as might have been expected. When GU samples were fit by GA-PDF's the indications given by the statistics δ and D_o regarding the parent PDF are, in general, erroneous (see Table 5.10b). Particularly for the Run 1GU5 ($\sigma_k^2 = 0.7311$), the statistics δ and D_o show that GU-data are fit

Table 5.8. Discrimination of PDF's - LN Data

(a) LN Data Fit to LN and GA Distributions by Least Squares (n = 100)

σ_y	σ_κ^2	No of Samples	No. of cases close fit indicated by						Rejections at 5% Level				Run No.
			Do		δ		SSE		χ^2 test		K-S test		
			LN	GA	LN	GA	LN	GA	LN	GA	LN	GA	
0.3	.0942	25	10	15	17	8	16	9	2	2	0	0	1LN5
0.4	.1735	25	17	8	21	4	18	6	0	7	0	0	2LN5
0.5	.2840	25	17	8	22	3	16	9	0	8	0	0	3LN4
0.7	.6323	22*	18	4	21	1	17	5	4	10	0	0	4LN4
	TOTAL	97	62	35	81	16	67	29	6	27	0	0	

* For 3 samples, the Sample Variance exceeded 1.0; hence, gamma fit not made.

(b) LN Data Fit to LN and GU Distributions by Least Squares (n = 100)

σ_y	σ_κ^2	No of Samples	No of cases close fit indicated by						Rejections at 5% Level				Run No.
			Do		δ		SSE		χ^2 test		K-S test		
			LN	GU	LN	GU	LN	GU	LN	GU	LN	GU	
0.3	.0942	25	19	6	15	10	15	10	2	2	0	0	1LN5
0.4	.1735	25	12	13	14	11	13	12	0	1	0	0	2LN5
0.5	.2840	25	16	9	19	6	16	9	0	3	0	0	3LN4
0.7	.6323	25	21	4	24	1	23	2	4	21	0	2	4LN4
	TOTAL	100	68	32	72	28	67	33	6	27	0	2	

(c) LN Data Fit to LN, GA and GU Distributions by Least Squares (n = 100)

σ_y	σ_κ^2	No of Samples	No of cases close fit indicated by									Run No
			Do			δ			SSE			
			LN	GA	GU	LN	GA	GU	LN	GA	GU	
0.3	.0942	25	4	15	6	8	8	9	6	9	10	1LN5
0.4	.1735	25	11	6	8	14	3	8	12	6	7	2LN5
0.5	.2840	25	16	4	5	19	2	4	15	5	5	3LN4
0.7	.6323	25	18	7	0	21	4	0	16	8	1	4LN4
	TOTAL	100	49	32	19	62	17	21	49	28	23	

Table 5.9. Discrimination of PDF's - GA Data

(a) GA Data Fit to LN and GA Distributions by LS (n = 100)

α, β σ_K^2	No of Samples	No of cases close fit indicated by						No of rejections at 5%				Run No.
		Do		δ		SSE		χ^2 test		K-S test		
		GA	LN	GA	LN	GA	LN	GA	LN	GA	LN	
3 .3333	25	20	5	16	9	16	9	3	4	0	1	1GA5
5 .2000	25	19	6	17	8	15	10	3	7	0	2	2GA5
7 .1429	25	22	3	17	8	18	7	3	6	0	0	3GA6
10 .1000	25	22	3	15	10	19	6	1	3	0	0	4GA5
TOTAL	100	83	17	65	35	68	32	10	20	0	1	

(b) GA Data Fit to GA and GU Distributions by LS (n = 100)

α, β σ_K^2	No of Samples	No of cases close fit indicated by						No of rejections at 5%				Run No.
		Do		δ		SSE		χ^2 test		K-S test		
		GA	GU	GA	GU	GA	GU	GA	GU	GA	GU	
3 .3333	25	14	11	17	8	16	9	3	5	0	0	1GA5
5 .2000	25	19	6	18	7	12	13	3	6	0	0	2GA5
7 .1429	25	20	5	15	10	15	10	3	4	0	0	3GA6
10 .1000	25	23	2	17	8	19	6	1	4	0	0	4GA5
TOTAL	100	76	24	67	33	62	38	10	19	0	0	

(c) Gamma Data Fit to LN, GA, and GU Distributions by LS (n = 100)

α, β σ_K^2	No of Samples	No of cases close fit indicated by									Run No.
		Do			δ			SSE			
		GA	LN	GU	GA	LN	GU	GA	LN	GU	
3 .3333	25	9	5	11	12	8	5	8	8	9	1GA5
5 .2000	25	16	6	3	14	6	5	9	8	8	2GA5
7 .1429	25	20	2	3	15	2	8	14	3	8	3GA6
10 .1000	25	22	1	2	15	7	3	18	2	5	4GA5
TOTAL	100	67	14	19	56	23	21	49	21	30	

Table 5.10. Discrimination of PDF's - GU Data

(a) GU Data Fit to GU and LN Distributions by LS (n = 100)

α	σ_κ^2	No of Samples	No of cases close fit indicated by Do		δ		SSE		Rejections at 5% level χ^2 test		K-S test		Run No.
			GU	LN	GU	LN	GU	LN	GU	LN	GU	LN	
1.5	.7311	25	22	3	18	7	19	6	6	9	1	7	1GU5
2.0	.4112	25	23	2	20	5	20	5	3	14	0	4	2GU4
3.0	.1828	25	17	8	13	12	14	11	2	2	0	0	3GU5
4.0	.1028	25	10	15	13	12	12	13	2	3	0	0	4GU5
	TOTAL	100	72	28	64	36	65	35	13	28	1	11	

(b) GU Data Fit to GU and GA Distributions by LS (n = 100)

α	σ_κ^2	No of Samples	No of cases close fit indicated by Do		δ		SSE		Rejections at 5% level χ^2 test		K-S test		Run No.
			GU	GA	GU	GA	GU	GA	GU	GA	GU	GA	
1.5	.7311	25	11	14	10	15	15	10	6	0	1	0	1GU5
2.0	.4112	25	18	7	16	9	20	5	3	5	0	0	2GU4
3.0	.1828	25	12	13	13	12	14	11	2	2	0	0	3GU5
4.0	.1028	25	13	12	19	6	18	7	2	5	0	0	4GU5
	TOTAL	100	54	46	58	42	67	33	13	12	1	0	

(c) GU Data Fit to FU, LN, and GA Distributions by LS (n = 100)

α	σ_κ^2	No of Samples	No of cases close fit indicated by Do			δ			SSE			Run No.
			GU	LN	GA	GU	LN	GA	GU	LN	GA	
1.5	.7311	25	11	0	14	9	1	15	15	1	9	1GU5
2.0	.4112	25	18	2	5	16	3	6	19	3	3	2GU4
3.0	.1828	25	7	7	11	3	10	12	6	10	9	3GU5
4.0	.1028	25	9	7	9	13	8	4	12	9	4	4GU5
	TOTAL	100	45	16	39	41	22	37	52	23	25	

to a GA-PDF better than the GU-PDF. The reason for such an occurrence is probably that the negative variates generated in Gumbel data (about 8% when $\delta_k^2 = 0.73$) have been discarded, and as a result, larger errors must have occurred towards K-S and chi-square test statistics when GU - data were fit to GU - PDF than when GU - data were fit to GA - PDF.

Closeness of LN-GU and GA-GU in certain ranges of σ_k^2 also hampered identification of parent PDF (by δ, D_o and SSE) when samples were fit to the three PDF's and compared simultaneously. LN and GA are satisfactorily identified only in a σ_k^2 range in which they are distinctly different from the other two PDF's (see Part c, of Tables 5.8 and 5.9). In case of GU samples, when fits to the three PDF's are compared simultaneously (see Table 5.10c), identification of the parent PDF (by δ, D_o and SSE) appears to be almost impossible for the range of σ_k^2 (0.12 to 0.73) investigated. The reason for such an occurrence may be the closeness of GU to LN or GA at low σ_k^2 (≤ 0.4) and use of only positively generated GU variates in GU samples.

For LN and GA samples (which are, in general, different from each other at any σ_k^2) the relative performance of δ and D_o as discriminators of PDF's is not consistent. While the statistic, δ, better identified the parent PDF (in 68% to 88% samples) when LN samples were fit to LN and GA (see Table 5.8a), the statistic, D_o, better identified the parent PDF (in case of 76% to 88% samples) when GA samples were fit to LN and GA (see Table 5.9a). In fact, for LN samples of Run 1LN5 ($\sigma_k^2 = 0.0942$, see Table 5.8a) GA was identified as the parent PDF for 60% of the samples by the statistics, D_o. (SSE is the least likely, 60% - 76%, of the three statistics to identify the parent PDF).

Tables 5.11 through 5.13 present the mean and variance (denoted by a bar and var, respectively) of δ, D_o and SSE, respectively, of the fits described in Tables 5.8 through 5.10.

Table 5.11 shows that, on the average, the chi-square statistics (i.e., $\overline{\delta}$LN, $\overline{\delta}$GA, or $\overline{\delta}$GU) has the highest value for the parent PDF when data were fitted to different PDF's. (The exception was GU data with $\sigma_k^2 = 0.7311$. The reason for this is that the modified GU data readily fit a GA at this variance). The above result shows that on the average, δ will identify the parent PDF. To examine how sensitive are the values of δ in identifying parent PDF, the ratios of $\overline{\delta}$LN/$\overline{\delta}$GA and $\overline{\delta}$GA/$\overline{\delta}$LN were evaluated for LN and GA data, respectively. The following table summarizes the two ratios

LN Data		GA Data	
σ_k^2	$\overline{\delta}$LN/$\overline{\delta}$GA	σ_k^2	$\overline{\delta}$GA/$\overline{\delta}$LN
.0942	1.21	.1000	1.21
.1735	1.73	.1429	1.21
.2840	1.98	.2000	1.50
.6323	2.32	.3333	1.28

The above table shows that the δ ratios are, in general, larger for LN data than for GA data. This could be the reason for better identification of LN by δ which was discussed earlier.

Table 5.12 shows that for LN and GA data, on the average, $\overline{D}_o$ is the smallest for the parent PDF when data were fit to LN, GA and GU

Table 5.11. PDF Discrimination by Chi-Square Statistic, δ

(n = 100, Number of Samples = 25)

Run No.	Data	σ_K^2	Data Fit by LS to LN		GA		GU	
			$\bar{\delta}$LN	Var δLN	$\bar{\delta}$GA	Var δGA	$\bar{\delta}$GU	Var δGU
1LN5	LN	.0942	.453	.080	.374	.067	.474	.094
2LN5		.1735	.442	.081	.255	.054	.400	.051
3LN4		.2840	.575	.073	.290	.080	.371	.078
4LN4		.6323*	.395	.105	.170	.059	.051	.023
4GA5	GA	.1000	.323	.105	.391	.080	.301	.089
3GA6		.1429	.368	.114	.444	.106	.411	.116
2GA5		.2000	.269	.118	.404	.114	.370	.110
1GA5		.3333	.282	.065	.362	.083	.258	.081
4GU5	GU	.1208	.378	.082	.270	.074	.403	.083
3GU5		.1828	.390	.070	.418	.097	.446	.073
2GU4		.4112	.151	.051	.300	.058	.326	.054
1GU5		.7311	.134	.030	.399	.067	.271	.054

$\bar{\delta}_{LN}$ = mean of (25) sample values of δ when data were fit to LN ($\bar{\delta}$ may be replaced by $\bar{D}$o or $\bar{SSE}$ and LN may be replaced by GA or GU to give the corresponding meaning)

* Three samples did not coverge for GA. $\bar{\delta}$ is the average of 22 values; the underlined values indicate the best PDF.

Table 5.12. PDF Discrimination by K-S Statistic, Do

(n = 100, Number of samples = 25)

Run No.	Data	σ_κ^2	Data fit by LS to LN		GA		GU	
			$\overline{Do}$,LN	Var Do,LN	$\overline{Do}$,GA	Var Do,GA	$\overline{Do}$,GU	Var Do,GU
1LN5	LN	.0942	.0597	$.263\times10^{-3}$	.0562	$.197\times10^{-3}$	.0661	$.361\times10^{-3}$
2LN5		.1735	.0546	$.080\times10^{-3}$	.0603	$.214\times10^{-3}$	.0570	$.139\times10^{-3}$
3LN4		.2840	.0560	$.146\times10^{-3}$	.0647	$.220\times10^{-3}$	.0635	$.320\times10^{-3}$
4LN4		.6323	.0588	$.381\times10^{-3}$	.0685	$.406\times10^{-3}$	.0955	$.771\times10^{-3}$
4GA4	GA	.1000	.0745	$.306\times10^{-3}$	.0610	$.161\times10^{-3}$	.0795	$.412\times10^{-3}$
3GA6		.1429	.0764	$.450\times10^{-3}$	.0594	$.305\times10^{-3}$	.0667	$.322\times10^{-3}$
2GA5		.2000	.0821	$.790\times10^{-3}$	.0646	$.475\times10^{-3}$	.0679	$.589\times10^{-3}$
1GA5		.3333	.0815	$.649\times10^{-3}$	.0574	$.187\times10^{-3}$	.0622	$.475\times10^{-3}$
4GU5	GU	.1208	.0534	$.225\times10^{-3}$	.0544	$.490\times10^{-3}$	.0555	$.202\times10^{-3}$
3GU5		.1828	.0668	$.256\times10^{-3}$	.0605	$.276\times10^{-3}$	.0612	$.299\times10^{-3}$
2GU4		.4112	.1040	$.113\times10^{-2}$	.0694	$.449\times10^{-3}$	.0621	$.420\times10^{-3}$
1GU5		.7311	.1213	.097	.0686	$.361\times10^{-3}$	.0809	$.936\times10^{-3}$

For a definition of $\overline{Do}$,LN etc., see Table 5.11.

The underlined values indicate the best PDF.

Table 5.13. PDF Discrimination by SSE

(n = 100, Number of Samples = 25)

Run No.	Data	σ_κ^2	Data Fit by LS to LN		GA		GU	
			$\overline{SSE}_{LN}$	Var SSE_{LN}	$\overline{SSE}_{GA}$	Var SSE_{GA}	$\overline{SSE}_{GU}$	Var SSE_{GU}
1LN5	LN	.0942	.00744*	$.283\times10^{-4}$	.00753	$.281\times10^{-4}$	.00743	$.292\times10^{-4}$
2LN5		.1735	.00657	$.153\times10^{-4}$	.00696	$.150\times10^{-4}$	.00674	$.014\times10^{-4}$
3LN4		.2840	.00490	$.066\times10^{-4}$	.00544	$.086\times10^{-4}$	.00529	$.085\times10^{-4}$
4LN4		.6323	.00626	$.209\times10^{-4}$	.00784	$.025\times10^{-4}$	.00849	$.267\times10^{-4}$
4GA5	GA	.1000	.00785	$.130\times10^{-4}$	.00744	$.120\times10^{-4}$	.00789	$.130\times10^{-4}$
3GA6		.1429	.00812	$.168\times10^{-4}$	.00779	$.137\times10^{-4}$	.00792	$.152\times10^{-4}$
2GA5		.2000	.00810	$.160\times10^{-4}$	.00791	$.165\times10^{-4}$	.00791	$.162\times10^{-4}$
1GA5		.3333	.00855	$.139\times10^{-4}$	.00828	$.181\times10^{-4}$	.00868	$.230\times10^{-4}$
4GU5	GU	.1208	.00789	$.050\times10^{-4}$	.00845	$.120\times10^{-4}$	.00779	$.095\times10^{-4}$
3GU5		.1828	.00694	$.123\times10^{-4}$	.00717	$.137\times10^{-4}$	.00692	$.014\times10^{-4}$
2GU4		.4112	.00969	$.111\times10^{-4}$	.00837	$.096\times10^{-4}$	.00787	$.087\times10^{-4}$
1GU5		.7311	.00903	$.176\times10^{-4}$	.00709	$.160\times10^{-4}$	.00664	$.116\times10^{-4}$

For a definition of $\overline{SSE}_{LN}$ etc., see Table 5.11.

* $\overline{SSE}_{LN}$ and $\overline{SSE}_{GU}$ differ by only .00001; hence, both are adjudged as equal.

The underlined values indicate the best PDF.

(the exception was LN data at $\sigma_k^2 = 0.0942$ at which LN and GA do not differ greatly, see Chapter III). The above result shows that, on the average, D_o identifies the parent PDF in case of LN and GA. The GU was not satisfactorily identified by D_o because of the overlapping nature of GU data with LN or GA at the appropriate σ_k^2. To examine the sensitivity of the values of D_o in identifying the parent PDF, the ratios of $\overline{D}_o$,GA/$\overline{D}_o$,LN and $\overline{D}_o$,LN/$\overline{D}_o$,GA were evaluated for LN and GA data, respectively. The following table summarizes the two ratios

LN Data		GA Data	
σ_k^2	$\overline{D}_o$,GA/$\overline{D}_o$,LN	σ_k^2	$\overline{D}_o$,LN/$\overline{D}_o$,GA
.0942	0.94	.1000	1.22
.1735	1.10	.1429	1.29
.2840	1.16	.2000	1.27
.6323	1.16	.3333	1.42

The above table shows that the D_o ratios are larger for GA data than for LN data. This could be the reason for better identification of GA by D_o which was discussed earlier.

Table 5.13 shows that, on the average, SSE has the smallest value for the parent PDF when data were fitted to different PDF's. However, an examination of the values of $\overline{SSE}$ shows that, in general, SSE is not sufficiently sensitive to serve as an effective identifier of PDF's which may be illustrated by the following table.

LN Data		GA Data	
σ_k^2	$\overline{SSE}_{GA}/\overline{SSE}_{LN}$	σ_k^2	$\overline{SSE}_{LN}/\overline{SSE}_{GA}$
.0942	1.01	.1000	1.06
.1735	1.06	.1429	1.04
.2840	1.11	.2000	1.02
.6323	1.25	.3333	1.03

Thus, the findings of Study No. 2 shows that, in general, SSE is not sufficiently sensitive to serve as an (effective) identifier of PDF's and although the statistics δ and D_o identify PDF's in most cases some inconsistencies (described above) occur in their performance as identifiers of PDF's.

The reasons why a certain statistic is better than others in identifying a PDF were not investigated in this study.

Chi-square and K-S statistics can be used as the basis for rejecting the null hypothesis, "There is nothing unusual about the data," i.e., the data fit the postulated distribution. This is the usual use of the statistics. The null hypothesis was tested at the 5% level of significance for each sample when fitted to each distribution, and the results are included in Tables 5.8 through 5.10. These results show that, in general, the K-S test, which is considered to be a more exact goodness-of-fit test, fails to reject a large number of incorrect hypotheses. Of the 600 cases in which the population differs from the hypothesized distribution, only 18 fits were rejected by the K-S test (at 5% level). However, the chi-square test rejected 133 of the

600 fits even though the chi-square test can, in some cases, give contradictory results depending on the grouping of the data (Benjamin and Cornell, 1970). The simulations thus found the three goodness-of-fit tests to be generally inadequate for identifying the correct parent distribution.

Study No. 3: PDF Discriminating Criterion Based on Tolerance Limits

For a given sample, one might intuitively feel that the tolerance limits (TF's) based on a PDF which is close to the population of the sample would show specific properties, such as they might be the lowest (or highest) of all such values computed compared to those determined from other PDF's. In an attempt to check whether the above expectation is true, three simulated data samples, one each from LN, GA and GU distributions were selected and the 90% upper tolerance limits were evaluated for the 100-year events by fitting each sample to LN, GA and GU by LS (Appendix A describes a method of constructing confidence regions and tolerance limits by LS) and the results compared.

Figures 5.4 through 5.6 show the histograms of the three simulated samples chosen. Figures 5.7 through 5.15 show the 90% confidence regions for the three samples by least squares fit to LN, GA and GU PDF's. These confidence regions show the wide range of values of the parameter estimates which remain within a certain confidence level. The shape of the confidence regions is elliptical, in general, and the major axis is approximately horizontal for the parameters of

lognormal and Gumbel distributions while it is inclined about 45° for the parameters of gamma distribution. Such an inclination in case of gamma distribution may be expected since gamma parameters are numerically equal for populations examined in this study.

To use confidence regions to construct an upper tolerance limit one determines the value of $\overline{K}_{\gamma u,\lambda}$ (see Appendix A) such that

$$\overline{K}_{\gamma u,\lambda} = \max_{\alpha,\beta \epsilon R_\lambda} \left[\overline{K}: \int_0^{\overline{K}} p(k;\alpha,\beta)dk = \gamma \right]$$

$1/(1-\gamma)$ gives the return period of $\overline{K}$ and λ represents the confidence level.

Thus the 90% upper tolerance limit of 100-year event is given by $K_{.99u,.9}$. While exact determination of $\overline{K}$ in the above equation is a formidable task, the calculation of $\overline{K}$ at a few critical points of the region R_λ will give an estimate of $\overline{K}_{\gamma u,\lambda}$ sufficient for most practical purposes. The portion of R_λ which needs to be examined to obtain $K_{\gamma u,\lambda}$ may be readily determined from the range of α and β (parameters of PDF) which maximize K_t, the random variable associated with the t-year event. (For example, parameters from the upper right quadrant of LN, lower tip of GA and the upper left quadrant of GU 90% confidence regions, see Figures 5.7 through 5.15, give larger K_t values).

The upper tolerance limits evaluated for the 100-year event at 90% confidence level for the nine cases described earlier are furnished in Table 5.14. Table 5.14 also shows the sample parameters when the three samples (one each from LN, GA and GU) were fit to the LN, GA and

Table 5.14. 90% Upper Tolerance Limits of 100-Year Event

(n = 100)

Run No.	Sample No.	Fitted PDF	Parameters of LS Fit		K_{S100} LS	Statistics of LS Fit		$\bar{K}_{0.99u,.90}$	$\frac{\bar{K}_{.99u,.90}}{\bar{K}_{S100}}$
			A	B		μ_F	σ_F^2		
4LN4	17	LN	-.2208	.6902	3.99	1.0175	.63	6.41	1.60
		GA	2.76	2.51	2.74	.909	.33	4.06	1.48
		GU	2.301	.5941	2.59	.845	.31	3.52	1.36
2GA5	8	LN	-.0899	.5185	3.05	1.046	.337	3.98	1.30
		GA	4.49	4.36	2.36	.970	.216	2.75	1.16
		GU	2.603	.7566	2.52	.978	.242	2.82	1.12
2GU4	7	LN	-.0186	.6767	4.74	1.234	.884	12.60	2.66
		GA	2.46	2.61	3.15	1.060	.431	4.46	1.46
		GU	1.989	.7137	3.03	1.004	.415	3.57	1.24

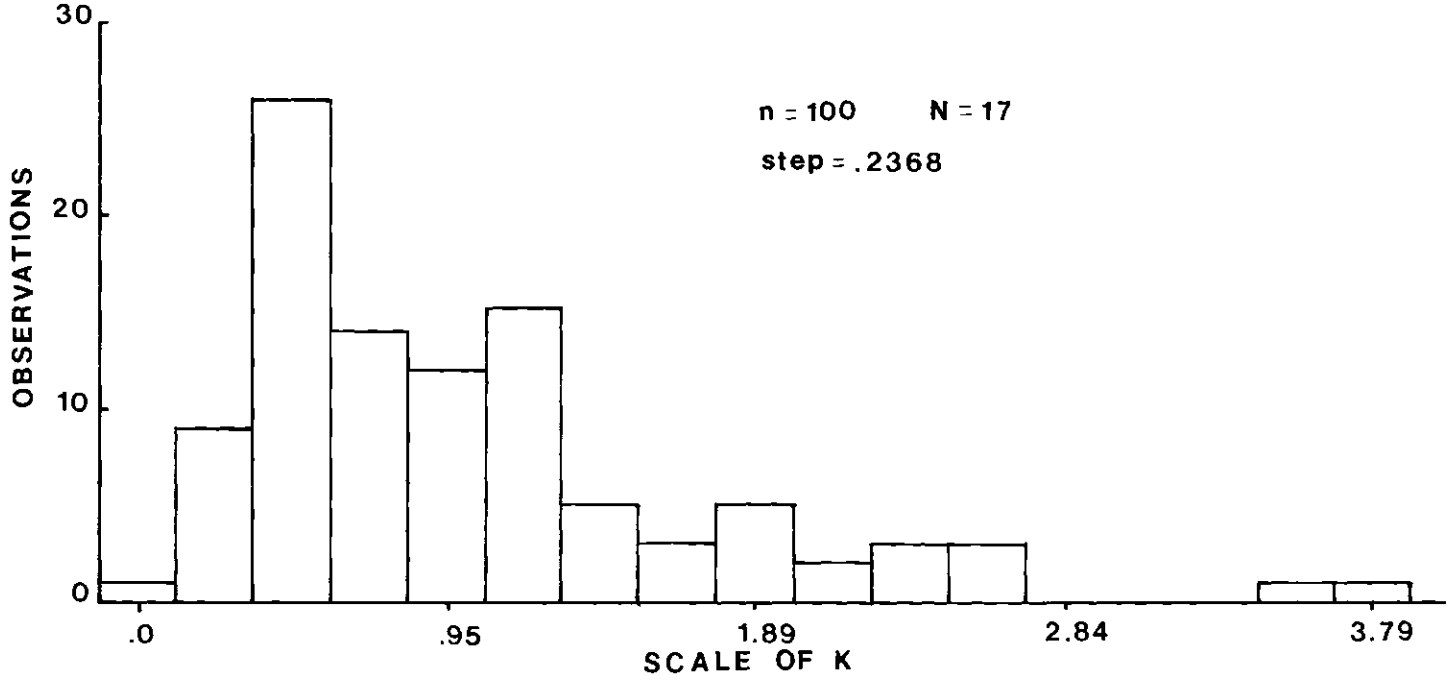


Figure 5.4 Histogram of a LN Sample (Sample No. 17, Run No. 4LN4)

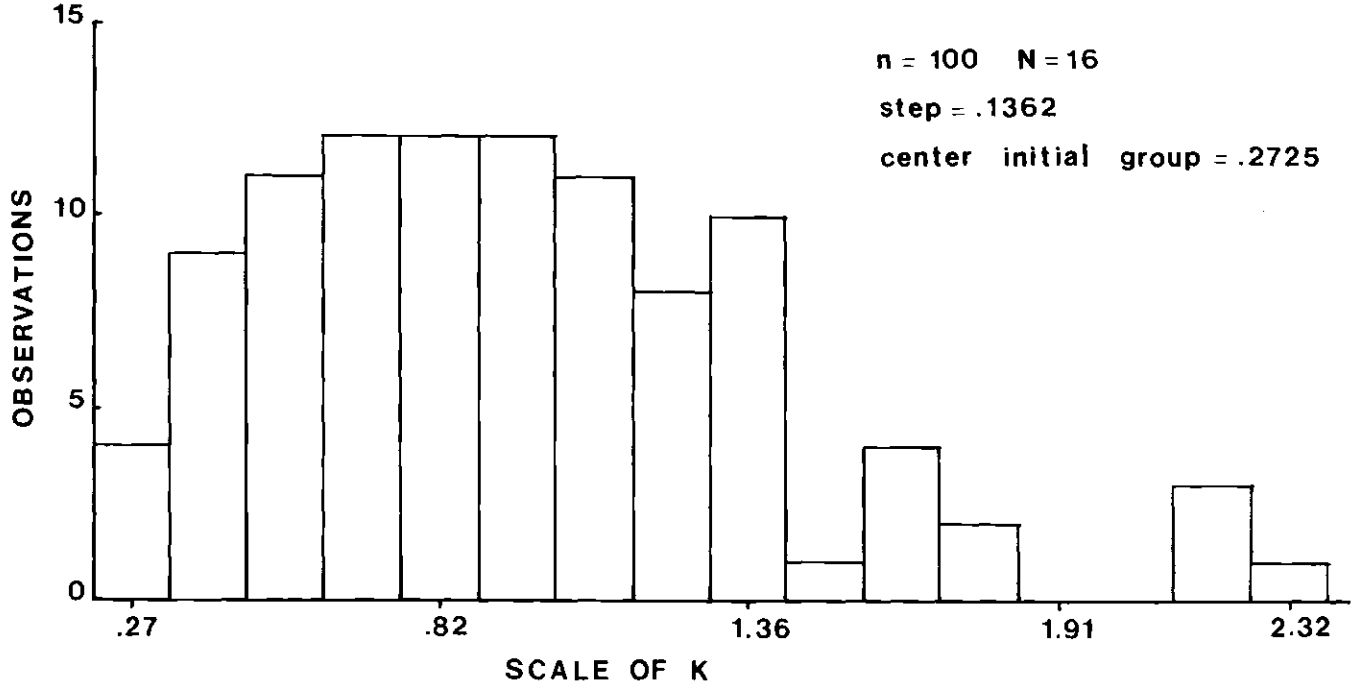


Figure 5.5 Histogram of a GA Sample (Sample No. 8, Run No. 2GA5)

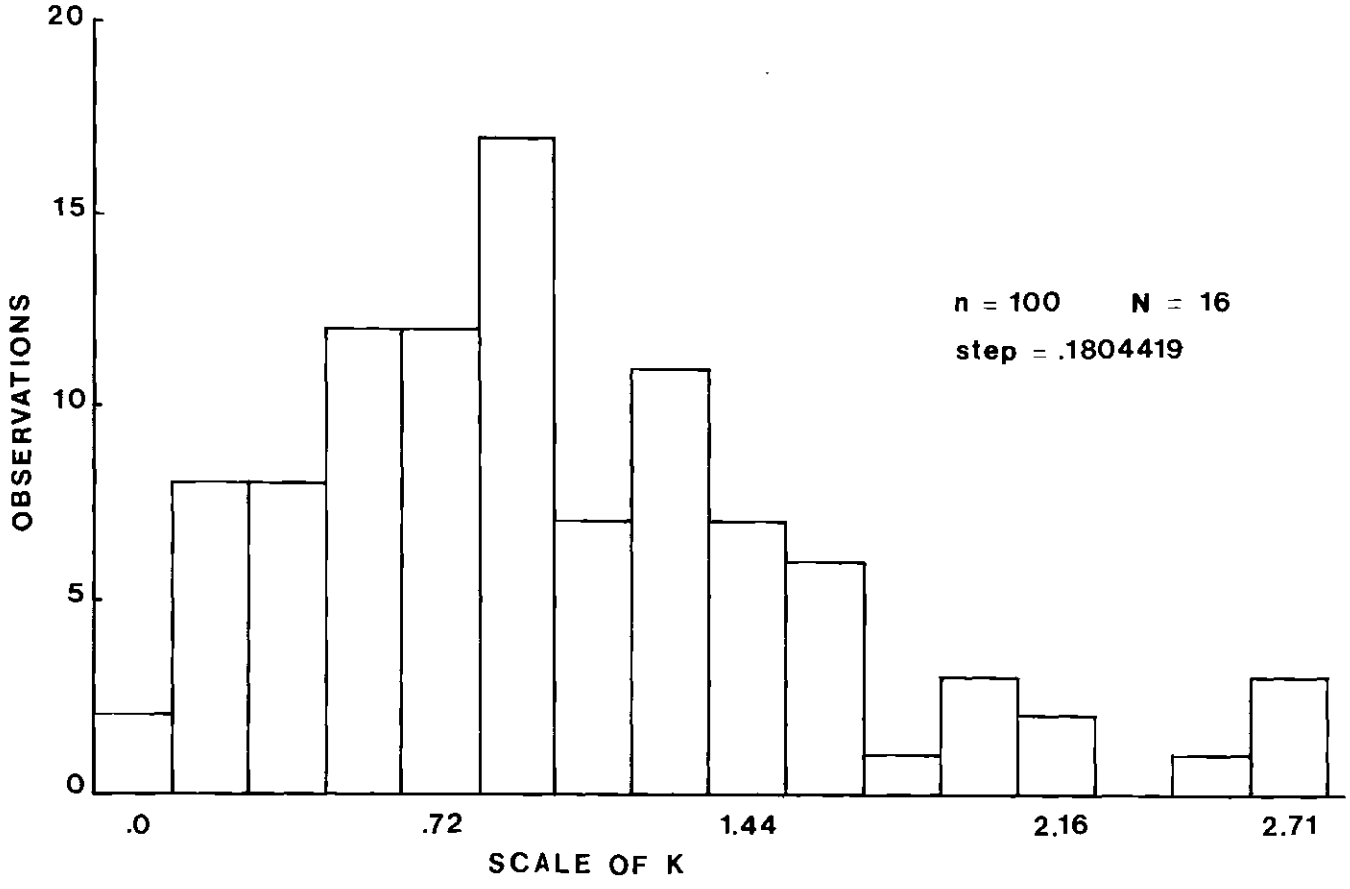


Figure 5.6 Histogram of a GU Sample (Sample No. 7, Run No. 2GU4)

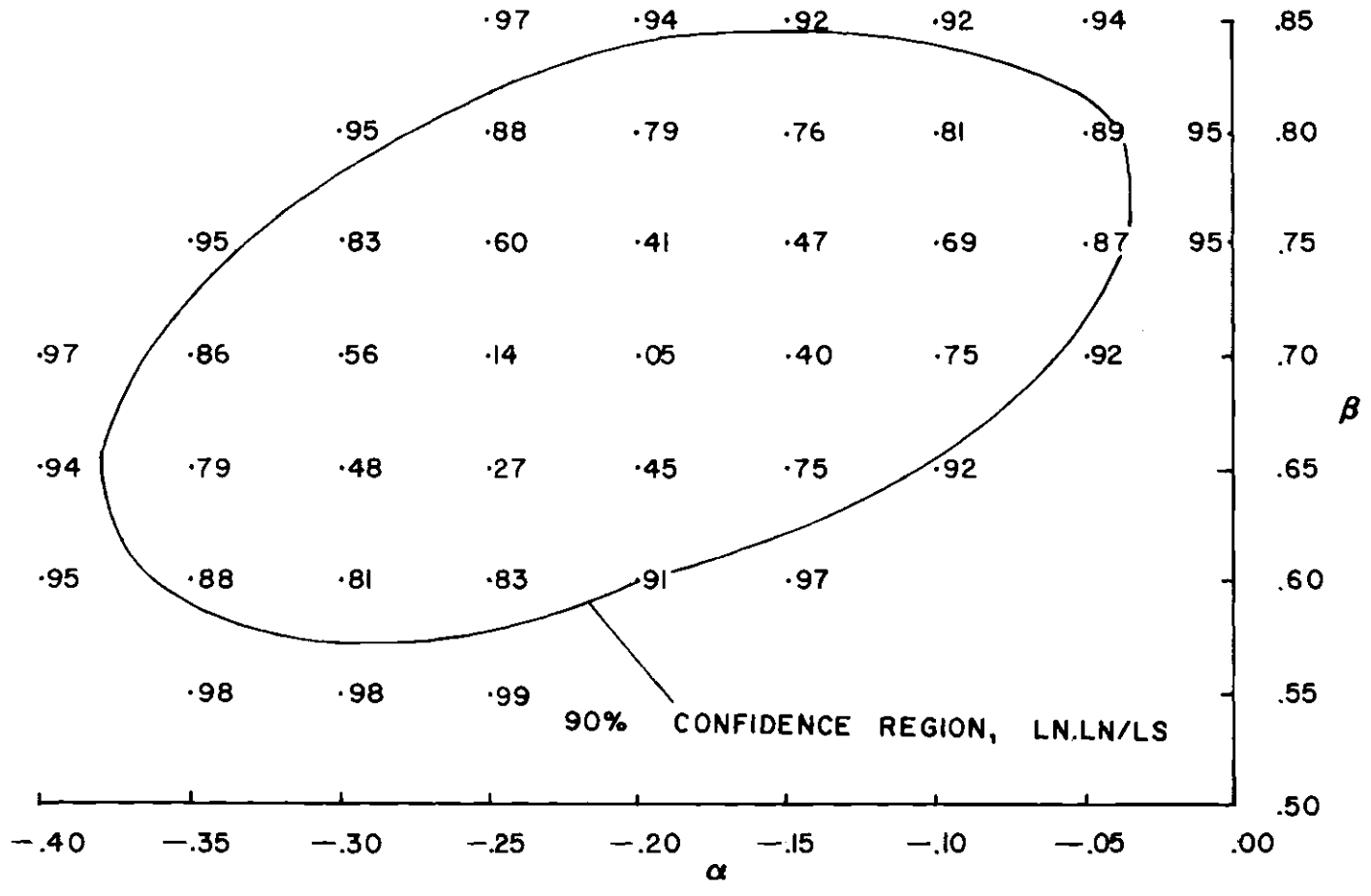


Figure 5.7 Confidence Region for a LN sample fit to LN PDF

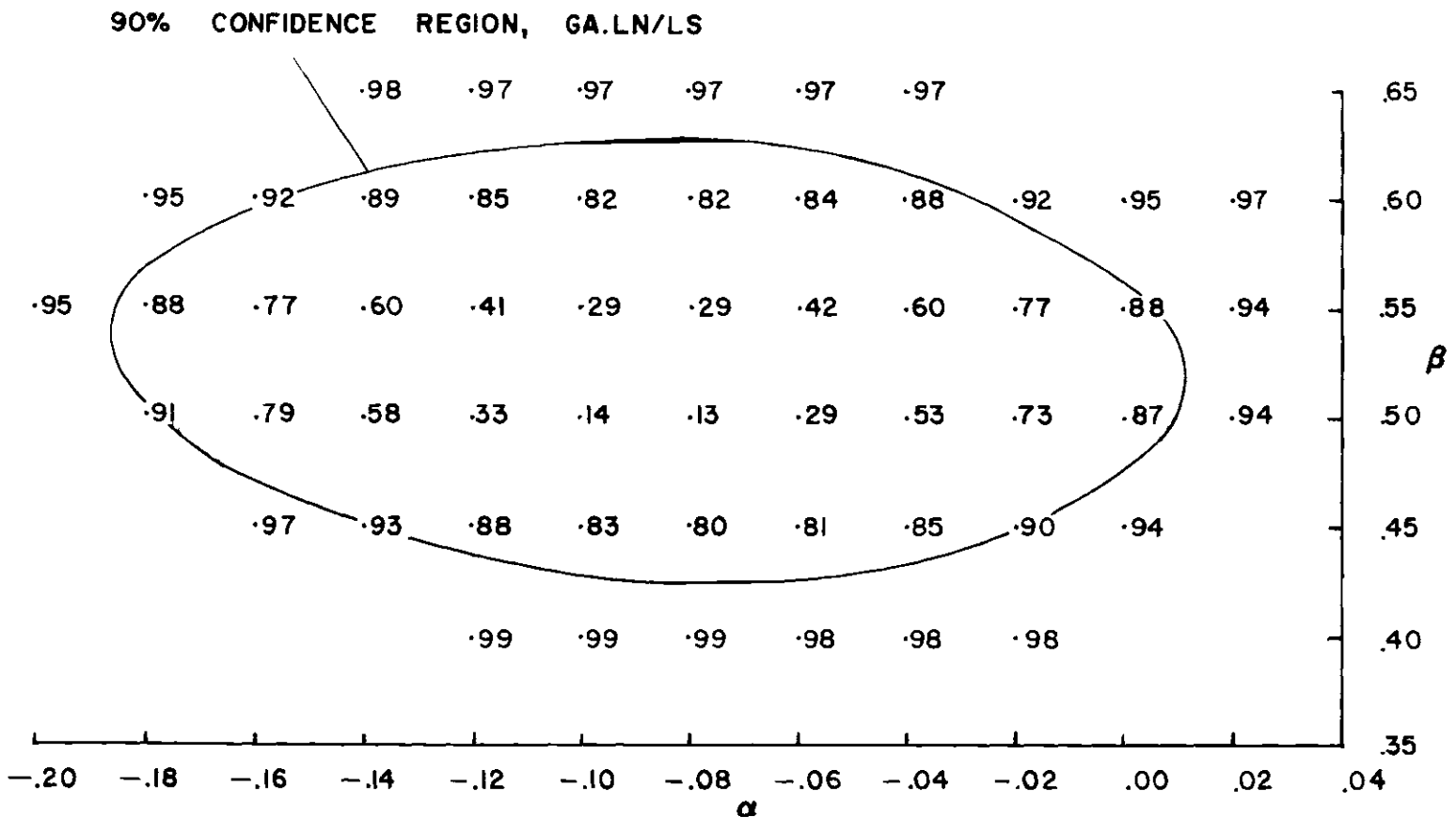


Figure 5.8 Confidence Region for a GA sample fit to LN PDF

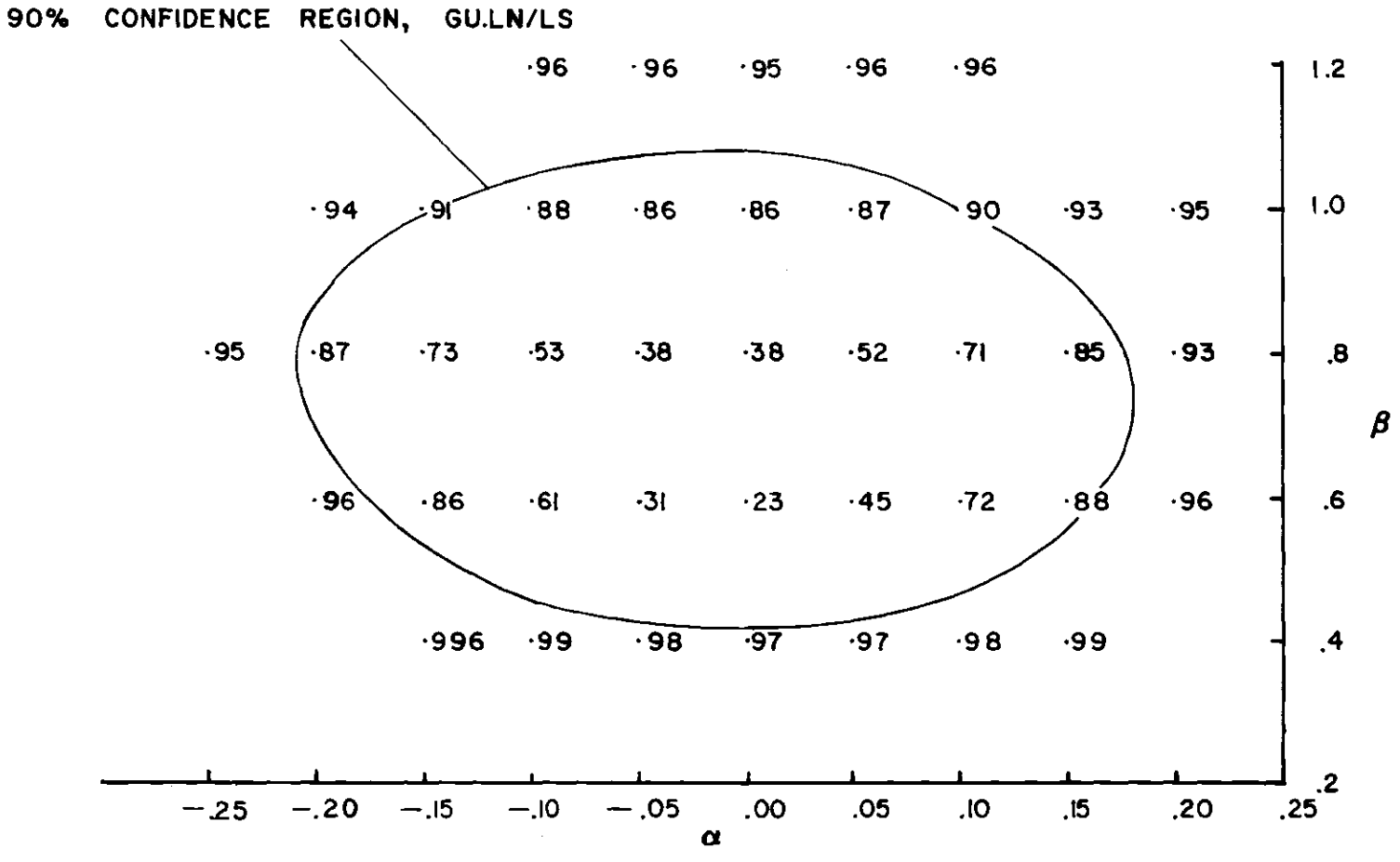


Figure 5.9 Confidence Region for a GU sample fit to LN PDF

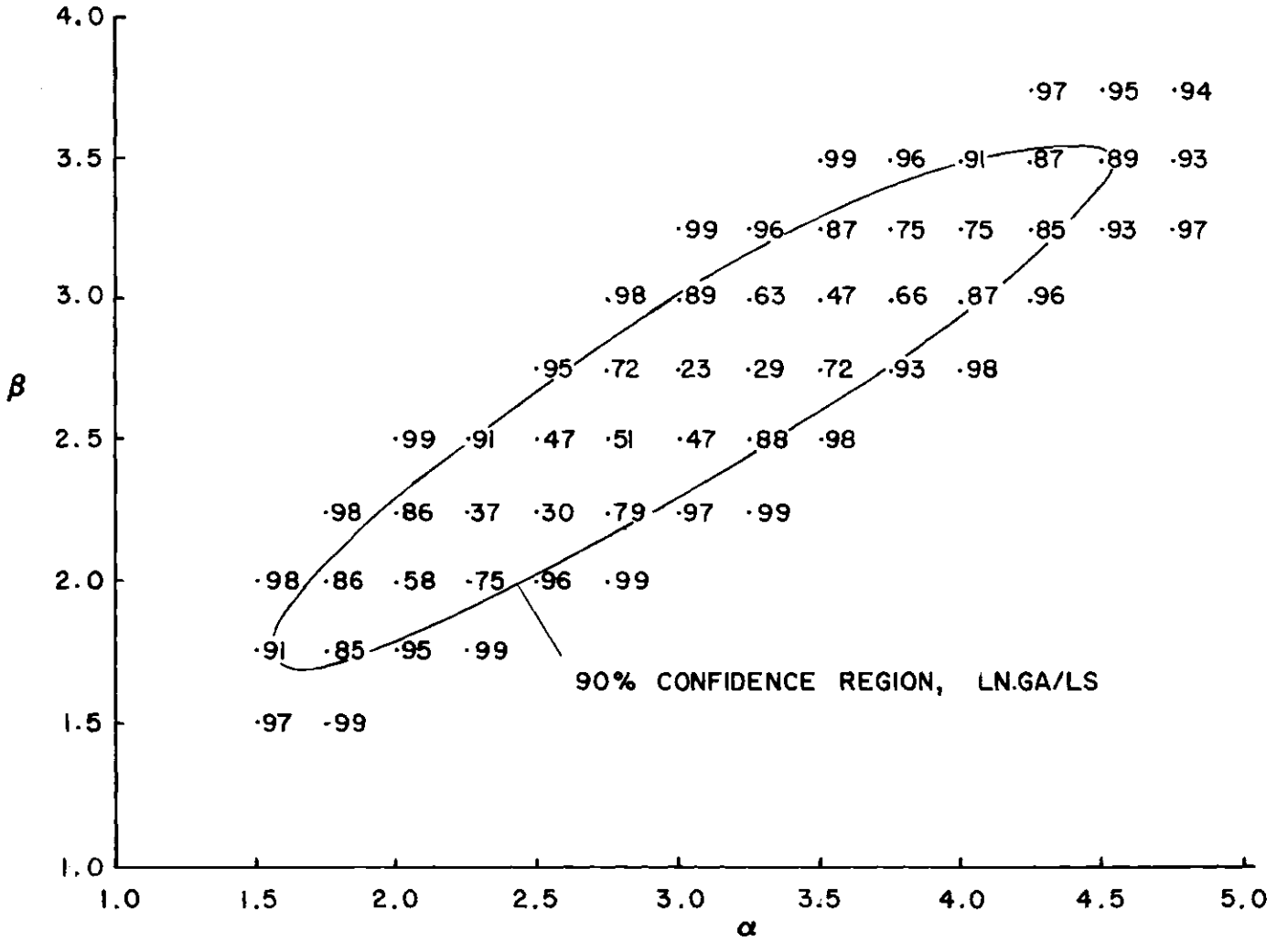


Figure 5.10 Confidence Region for a LN sample fit to GA PDF

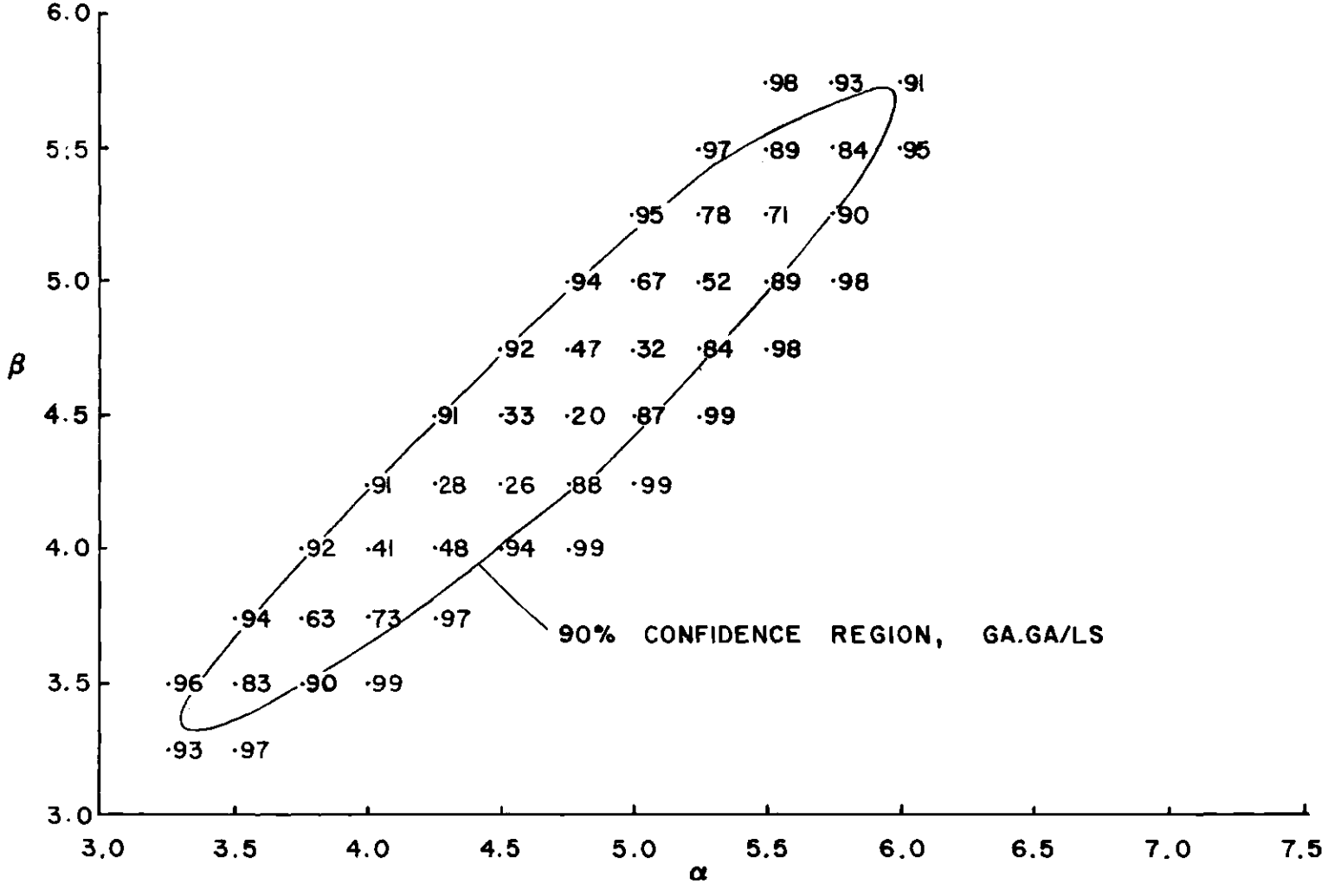


Figure 5.11 Confidence Region for a GA sample fit to GA PDF

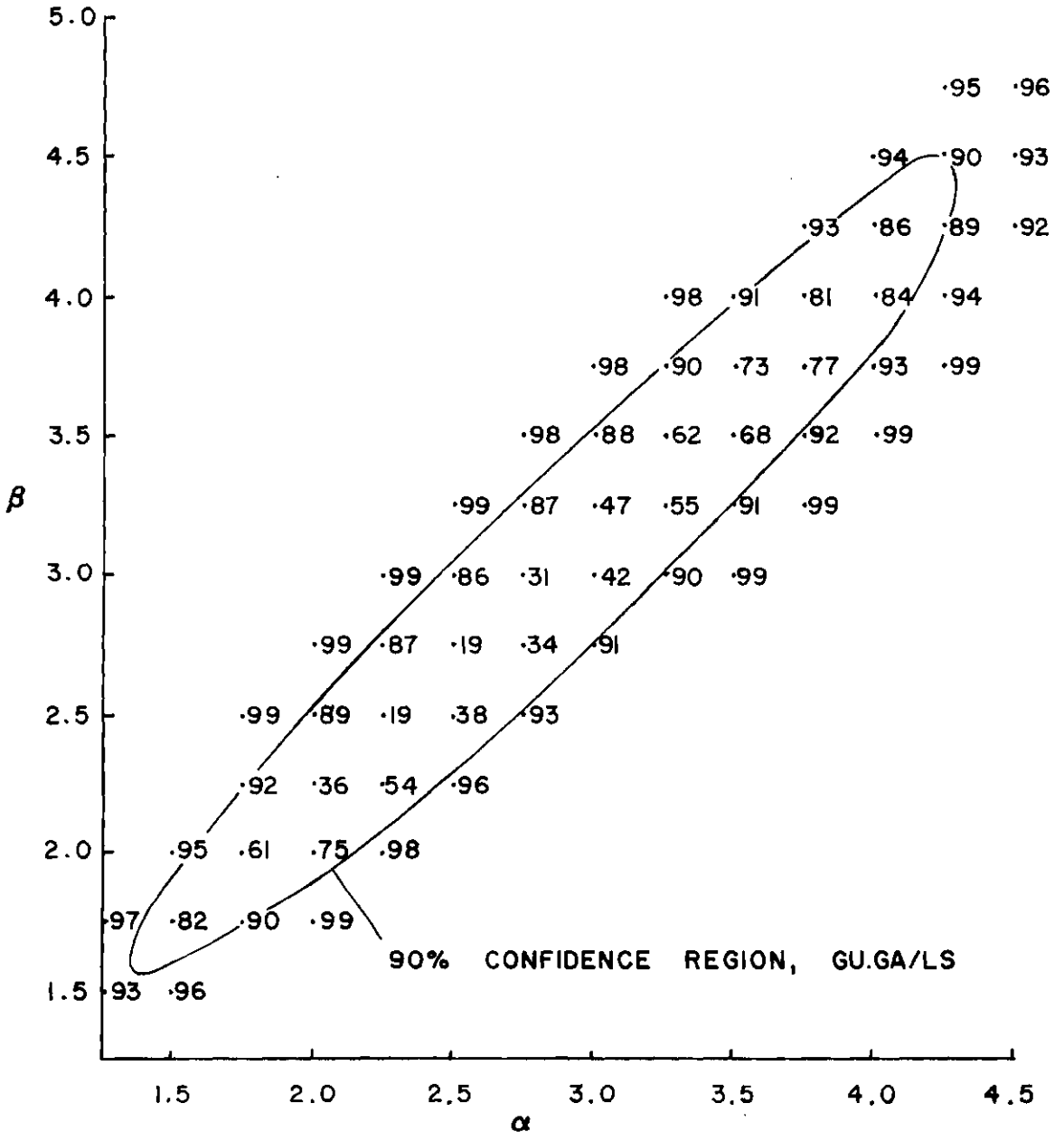


Figure 5.12 Confidence Region for a GU sample fit to GA PDF

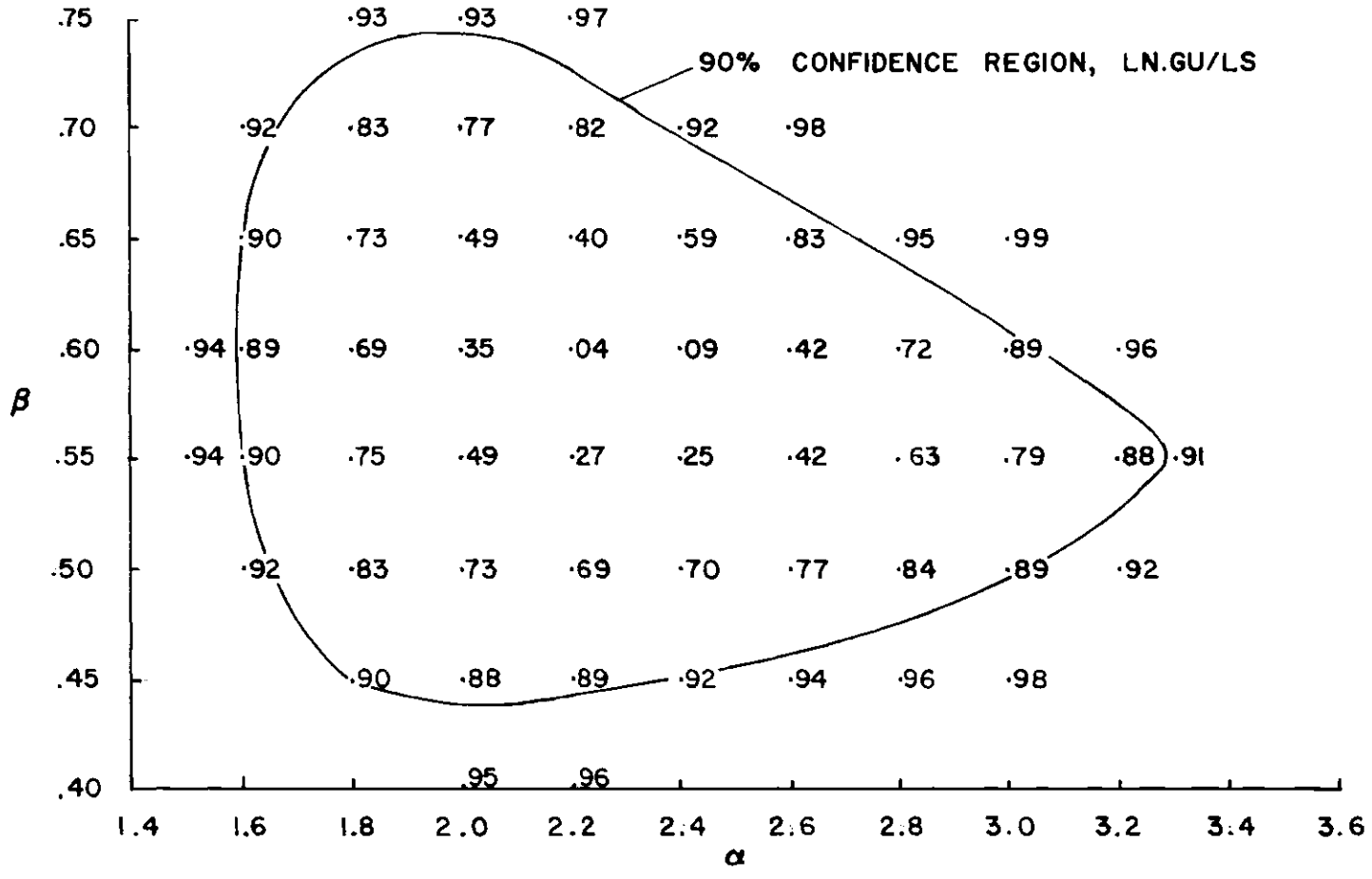


Figure 5.13 Confidence Region for a LN sample fit to GU PDF

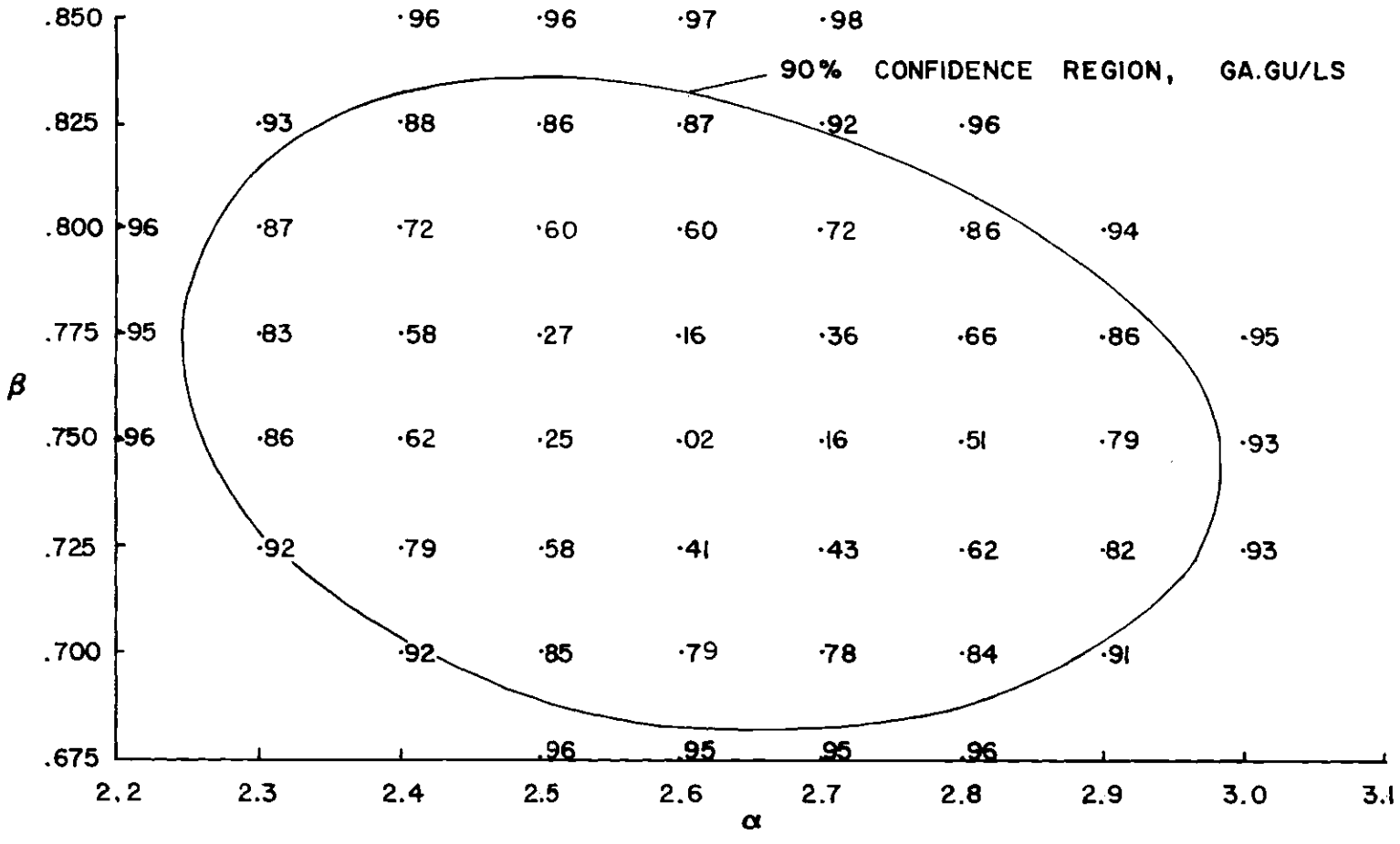


Figure 5.14 Confidence Region for a GA sample fit to GU PDF

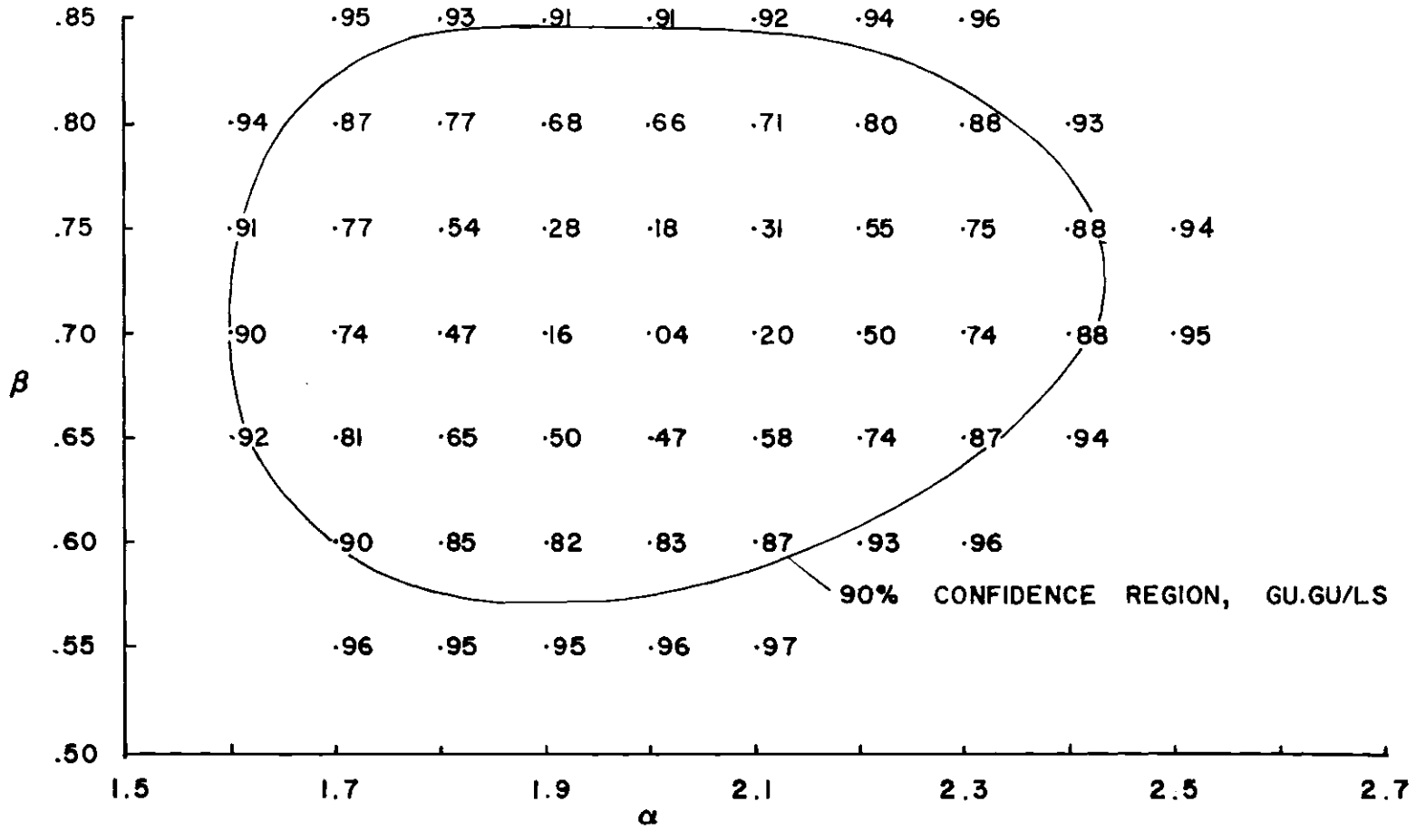


Figure 5.15 Confidence Region for a GU sample fit to GU PDF

GU PDF's by LS together with the corresponding 100-year sample (LS) prediction K_{S100}.

Table 5.14 shows that the TL's for a given sample based on a fit to the parent PDF do not show any unique properties. Like predictions (K_{S100}), the values of TL's are also found to be a function of the PDF to which the data sample was fit. In general, the magnitudes of TL's are more related to the predictions (and thus to the PDF) rather than the best (i.e., parent) PDF. The ratios $\overline{K}_{0.99u,.9}/K_{S100}$ (see Table 5.14) of parent fit also did not show any specific properties compared to those of other fits of a sample. Since the computations are tedious, no attempt has been made in this study to evaluate properties of TL's from samples at different levels of S_k^2 for each of the populations. From the results of Study No. 3, it may be tentatively stated that the TL's based on the parent PDF do not exhibit any unique properties and thus the TL's are not useful in identifying the parent PDF.

Summary

The three studies presented in this chapter show that, of the five PDF - discriminating criteria discussed in Chapter IV, the criterion based on TL's is not useful to identify PDF's. The three criteria based on the statistics, δ, D_o and SSE, on the average, identified the parent PDF. However, while these three statistics represent errors between the fitted PDF and the sample, the criterion given by Equation 5.2 includes important statistical parameters of the

sample and fitted PDF, i.e., S_k^2 and σ_F^2. Also, simulation showed that SSE was not sufficiently sensitive to discriminate PDF's and δ and D_o had a propensity to better identify one PDF than the other. Table 5.15 summarizes the relative performance of δ, D_o and σ_F^2/S_k^2 (or S_k^2/σ_F^2) in discriminating PDF's. In Table 5.15 only LN and GA results are included because these two PDF's are always different and thus afford better comparison of results. Table 5.15 shows that the value of the variance ratio is always larger than 1.0 for a PDF other than parent PDF (in Table 5.15, the variance ratio is so formulated as to obtain a value larger than 1.0 when the PDF differed from the parent PDF) and increased as σ_k^2 increased. Compared to the other discriminating statistics, the value of the variance ratio seems to be less dependent on the type of PDF from which the sample was drawn. Moreover, evaluation of variance ratio is computationally simpler than the evaluation of δ, D_o or SSE. Thus, the criterion given by Equation 5.2 was found to be the most useful in identifying the parent PDF of a sample. This criterion was applied to 67 real hydrologic samples and the results are presented in Chapter VI.

The 'Best Fit' criterion proposed in this study was based on the tacit assumptions that the variance of a sample approximately represents the population variance and the shape of the sample (distribution) approximately represents the shape of population distribution. The above assumptions are, at best, only statistical expectations. For some of the synthetic samples generated in this work the sample variance was found to be markedly different from the population

Table 5.15. PDF Discrimination by Different Statistics

(a) LN Data Fit to LN and GA

σ_κ^2	$\left(\frac{\bar{\delta}_{LN}}{\bar{\delta}_{GA}}\right)_{LS}$	$\left(\frac{\overline{Do}_{GA}}{\overline{Do}_{LN}}\right)_{LS}$	$\left(\frac{\overline{SSE}_{GA}}{\overline{SSE}_{LN}}\right)_{LN}$	$*\bar{S}_\kappa^2/\sigma_{F,GA/LS}^2$	$*\bar{S}_\kappa^2/\sigma_{F,\ GA/ML}^2$	$*\bar{S}_\kappa^2/\sigma_{F,GA/MCS}^2$
.094	1.21	0.94	1.01	1.20	1.09	1.08
.173	1.73	1.10	1.06	1.69	1.18	1.15
.284	1.98	1.16	1.11	1.73	1.28	1.23
.632	2.32	1.16	1.25	2.18	1.53	1.51

(b) GA Data Fit to GA and LN

σ_κ^2	$\left(\frac{\bar{\delta}_{GA}}{\bar{\delta}_{LN}}\right)_{LS}$	$\left(\frac{\overline{Do}_{LN}}{\overline{Do}_{GA}}\right)_{LS}$	$\left(\frac{\overline{SSE}_{LN}}{\overline{SSE}_{GA}}\right)_{LS}$	$*\sigma_{F,LN/LS}^2/\bar{S}_\kappa^2$	$\frac{*\sigma_{F,LN/ML}^2}{\bar{S}_\kappa^2}$	$*\bar{S}_\kappa^2/\sigma_{F,GA/MCS}^2$
.100	1.21	1.22	1.06	1.16	1.12	1.29
.143	1.21	1.29	1.04	1.31	1.17	1.45
.200	1.50	1.27	1.02	1.48	1.31	1.77
.333	1.28	1.42	1.03	1.95	1.59	2.23

* These ratios are approximately equal to 1.0 for the parent PDF's.

variance. For some exceptional lognormal samples (of size 100) the sample variance was found to be as low as 0.5 and as high as 2.9 times the population variance. The reasons for such occurrences and questions such as how the 'Best Fit' criterion would react when the criterion is applied to the exceptional samples mentioned above need to be investigated.

CHAPTER VI

"BEST FIT" CRITERION APPLIED TO REAL DATA

This chapter uses real data to examine the applicability of the "Best Fit" criterion (Equation 5.2) postulated in Chapter IV and corroborated by the results of simulation experiments in Study No. 1 of Chapter V. A second objective is to determine the effect of outliers on the results given by shape fitting methods like maximum likelihood and least squares.

Source of Data

The data used consists of the annual flood peaks for 67 streams throughout the United States as published in the U.S. Geological Survey Water Supply papers. For the compilation by Robey and Wallace (1969), the streamflow records were included if the record through 1960 was at least 49 years in length, the location of the gage had not changed appreciably within the period of record, the stream was unregulated, and the record had no gap of more than three years within the period of record. For this study, the streamflow records were updated to include the period from 1961 through 1970 (1973 for some stations). Appendix H gives a complete list of the stations used.

The 67 stream gauging stations are numbered in ascending order of sample variance, S_k^2 in Table 6.1 (see also Appendix H). Such

Table 6.1. Some Characteristics of Stream Gauging Stations Selected for Study,

STN NO (1)	S_K^2 (2)	n YEARS (3)	D.A. SQ.MI. (4)	$\bar{Q}$ CFS (5)	K_{max} (6)	% VARIANCE DUE TO K_{max} (7)	K_{min} (8)
1	0.0585	56	420.	1334.	1.61	12	0.50
2	0.0817	64	604.	3732.	1.83	13	0.39
3	0.0821	58	812.	2495.	1.76	12	0.61
4	0.0901	63	408.	10446.	1.76	10	0.41
5	0.0929	57	295.	5685.	2.18	27	0.54
6	0.1141	53	1690.	25794.	1.90	11	0.40
7	0.1259	66	163.	1899.	2.20	15	0.32
8	0.1307	62	1460.	9199.	2.07	14	0.38
9	0.1316	60	50.	398.	2.13	16	0.31
10	0.1329	64	872.	14156.	2.03	13	0.34
11	0.1353	77	1357.	35612.	2.18	13	0.32
12	0.1360	73	11220.	137078.	2.65	28	0.33
13	0.1455	75	1430.	3656.	1.88	7	0.31
14	0.1512	56	491.	14040.	2.28	20	0.43
15	0.1655	61	321.	7733.	2.33	18	0.39
16	0.1754	59	129.	667.	1.87	7	0.06
17	0.1861	59	559.	10797.	2.37	17	0.26
18	0.1880	60	300.	8408.	2.50	20	0.34
19	0.2120	77	2084.	40000.	2.77	19	0.29
20	0.2241	61	528.	12748.	2.76	23	0.40
21	0.2243	60	208.	7765.	2.58	19	0.30
22	0.2348	54	22.	1159.	2.33	14	0.30
23	0.2395	71	297.	3804.	2.59	15	0.27
24	0.2448	67	3583.	102563.	2.58	15	0.32
25	0.2518	63	22.	55.	2.77	20	0.19
26	0.2567	60	217.	907.	2.26	10	0.25
27	0.2766	62	657.	28298.	3.07	25	0.23
28	0.2775	64	3076.	70853.	3.29	30	0.26
29	0.2781	56	2140.	15811.	3.34	30	0.40
30	0.2943	69	6600.	71663.	2.61	13	0.32
31	0.3053	80	4400.	37897.	2.55	18	0.23
32	0.3055	67	1170.	19059.	2.89	18	0.41
33	0.3103	82	605.	13417.	2.74	12	0.33
34	0.3196	61	2400.	15013.	2.32	9	0.15
35	0.3329	59	33.	260.	3.09	21	0.16

Table 6.1. Some Characteristics of Stream Gauging Stations Selected for Study, continued

STN NO (1)	S_K^2 (2)	n YEARS (3)	D.A. SQ.MI. (4)	$\bar{Q}$ CFS (5)	K_{max} (6)	% VARIANCE DUE TO K_{max} (7)	K_{min} (8)
36	0.3359	62	678.	3516.	3.90	41	0.31
37	0.3487	66	7323.	48450.	3.30	23	0.24
38	0.3530	63	898.	11912.	2.64	12	0.12
39	0.3533	66	3495.	21389.	2.94	16	0.18
40	0.3557	76	388.	8586.	3.07	16	0.15
41	0.3750	81	4490.	55950.	3.47	20	0.18
42	0.3837	72	1230.	6086.	4.03	34	0.26
43	0.3850	67	15619.	41843.	2.89	14	0.17
44	0.4007	72	9651.	120393.	3.99	31	0.30
45	0.4032	56	98.	4338.	2.99	16	0.18
46	0.4044	61	171.	3940.	4.29	45	0.31
47	0.4048	63	6510.	27826.	2.62	10	0.12
48	0.4116	69	913.	18076.	3.41	21	0.19
49	0.4315	104	36800.	42544.	4.02	21	0.18
50	0.4704	59	5511.	16500.	3.48	22	0.09
51	0.4851	63	1090.	15301.	4.37	39	0.24
52	0.5009	56	510.	12479.	4.44	43	0.32
53	0.5075	54	3000.	3571.	4.51	45	0.30
54	0.5182	58	31.	1116.	6.00	95	0.44
55	0.5462	64	244.	9220.	3.07	12	0.09
56	0.5548	65	1599.	32864.	4.26	29	0.13
57	0.5851	66	3637.	17363.	4.23	27	0.15
58	0.5974	63	20.	33094.	3.96	22	0.24
59	0.6995	57	18.	32.	4.84	38	0.06
60	0.7043	62	934.	4991.	4.57	29	0.17
61	0.7063	52	321.	5925.	3.95	24	0.25
62	0.7775	52	74.	1034.	4.76	36	0.19
63	0.8681	53	46.	1635.	4.07	21	0.29
64	0.9613	62	53.	3448.	5.77	39	0.12
65	0.9738	56	34.	1943.	5.30	35	0.15
66	1.4175	64	253.	4463.	5.71	25	0.03
67	2.4943	55	16.	1175.	7.34	30	0.03

characteristics of stream gauge stations as the number of years of record (n), the drainage area (D.A.), and the mean annual peak discharge ($\bar{Q}$) are recorded in Table 6.1 with the magnitude of the sample variance, S_k^2, the highest recorded flow ($K_{max} = Q_{max}/\bar{Q}$) and the percent of the sample variance due to K_{max} for each station. The last mentioned quantity is particularly useful in analyzing outliers, and is defined as:

$$\text{percent variance due to } K_{max} = \left[\frac{(K_{max} - \bar{K})^2}{n - 1}/S_k^2\right] \times 100 \qquad (6.1)$$

Table 6.1 shows about 75% of the stations to have a sample variance, S_k^2, less than 0.5 and only two stations with S_k^2 larger than 1.0. The highest recorded annual peak flows at the 67 stations ranged from 1.6 to 7.3 times the mean annual peak.

Variance of Fitted PDF by the Shape Fitting Methods

Chapter V described how comparison of sample moments with the moments of a PDF fitted by a shape fitting method provides a criterion to identify the parent PDF. The "Best Fit" criterion is given by Equation 5.1.

The streamflow data for the 67 stations were fit to LN, GA and GU distributions by ML, LS and MCS methods and the ratio of variance of the fitted PDF (σ_F^2) to the sample variance was evaluated for each fit. Table 6.2 presents the variance ratios σ_F^2/S_k^2 resulting from each shape fitting method for each of the 67 stations. From these results the

samples may be divided into the three categories by the following procedure:

a) The LS and MCS methods were not viewed as two separate methods, but viewed as the two extreme cases of 'Weighted Least Squares' (WLS) method and WLS was used (see Section VI, Appendix A for justification of such a procedure. Section VII also describes the application of WLS in parameter estimation and in discriminating PDF's). The WLS does not give an explicit solution, but, in this method a ϕ value is chosen in the range of 0. to 1. to make the solution unbiased when the chosen PDF is the best. For the samples for which the applicable values of ϕ are 0. and 1. the LS and MCS solutions, respectively, will become solutions by WLS. Hence, the variance ratios by LS and MCS were included in Table 6.2. For cases for which the applicable value of ϕ lies between 0. and 1.0 for WLS solution, the variance ratio by WLS becomes unity; station numbers 2(GA), 3(LN,GU), 4(GA), 5(LN,GU), 7(LN,GU) are some such examples (for these PDF's the WLS method would give a variance ratio of unity for some as yet undetermined value of ϕ.)

b) To choose the best PDF both WLS and ML are equally applicable*. In case of a conflict, however, ML results may be generally

---

* Since ML and WLS are computationally two different methods, the choices of the 'Best Fit' by these methods are shown separately first (cols (10) and (11), Table 6.2) and then the two results are combined (col (12), Table 6.2). ML does not include GU.

preferred since WLS groups the data by which some vital information regarding the sample may be lost. In Table 6.2, ML and WLS gave conflicting results for 6 samples which were station numbers 21, 24, 27, 35, 37 and 49.

Category 1: The variance ratio, σ_F^2/S_k^2, is approximately unity (deviation not more than ± 0.15 from unity) for at least one of the hypothesized distributions. These samples may be called "Easy-to-Fit" samples because a parent PDF is identified by the "Best Fit" criterion. 50 of 67 samples of Table 6.2 (marked x in remarks column) are so categorized.

Category 2: The variance ratio is < 0.9 for LN and GA by ML, and most cases for all three hypothesized PDF's by WLS. 19 of 67 samples of Table 6.2 (marked "y" in remarks column) are found to be in this category. An examination of these samples revealed that they contained one or more outliers (data items far removed from the trend of the others). Hence these are called "the samples with outliers." (6 samples were found to be common for Categories 1 and 2).

Category 3: "Hard-to-Fit" samples. Four samples did not fit into the above two categories and are marked "z" in Table 6.2.

These three categories are discussed in the following pages.

"Easy to Fit" Samples

There are 50 samples for which the variance of fitted PDF by at least one of the shape fitting methods is reasonably close to the sample variance for some hypothesized PDF. The "Best Fit" criterion of Equation 5.2 indicates for these stations the PDF which best fits the

data. In determining the best PDF, preference was given to the variance ratio obtained by ML method since ML method is considered, in some respects, to be a statistically superior method (see Chapter II) and offers a solution in which the probability of occurrence of sample observations is maximized (see Chapter II).

The LS method, however, also proved to be a good identifier of the best PDF. Table 6.2 shows that in most cases for which the best PDF is identified the variance ratio evaluated by ML lies somewhere between the variance ratios evaluated by LS (ϕ = 0.0) and MCS (ϕ = 1.0) for the best PDF.

Six samples were found to be common for categories 1 and 2 because their variance ratios were less than 0.9 for LN and GA by ML but greater than 0.85 for one of the PDF's by one of the methods. These samples were included for discussion in this section and also in the next section.

The Gumbel distribution was never found to be unequivocally superior to both gamma and lognormal because, within the range of application of the Gumbel, it closely resembles the LN or the gamma (see Chapter III). For three samples at very low variance (S_k^2 = 0.08 to 0.13) all three PDF's were found to be equally suitable and for most samples with S_k^2 less than 0.2 the variance of the fitted PDF's differed by less than 10%. These results may be expected because the three PDF's closely resemble each other in this variance range (see Chapter III). Since the predictions (up to t = 500 years) given by the three PDF's also do not vary much in this variance range, the particular PDF

Table 6.2. Ratios of Variance of Fitted PDF to Sample Variance, (σ_F^2/S_k^2)

STN NO (1)	LOGNORMAL PDF ML (2)	LS (3)	MCS (4)	GAMMA PDF ML (5)	LS (6)	MCS (7)	GUMBEL PDF LS (8)	MCS (9)	PROBABLE BEST FIT BY ML (10)	WLS (11)	ML&WLS (12)	REMARKS (13)
1	1.021	0.946	00	1.085	0.853	11	1.180	00	LN	LN	LN	X
2	1.093	1.025	1.182	1.029	0.931	1.022	1.135	1.358	GA	GA	GA	X
3	0.921	0.782	1.032	0.945	0.680	0.955	0.845	1.009	GA	LN,GU	ALL	X
4	1.060	1.041	1.350	1.004	0.889	1.124	1.205	1.387	GA	GA	GA	X
5	0.876	1.014	0.928	0.894	0.880	0.899	1.108	0.935	GA	LN,GU	ALL	X,Y(27)
6	1.078	1.409	1.254	0.986	1.156	1.053	1.331	1.204	GA	GA	GA	X
7	1.049	0.799	1.253	0.948	0.679	1.028	0.883	1.207	GA	ALL	ALL	X
8	1.226	1.036	1.330	0.983	0.867	1.049	1.105	1.251	GA	GA	GA	X
9	0.988	0.636	1.161	0.908	0.561	0.980	0.716	1.112	LN	LN,GU	LN	X
10	1.205	1.075	1.322	1.025	0.907	1.033	1.097	1.244	GA	GA	GA	X
11	1.286	1.641	1.405	1.064	1.251	1.071	1.551	1.311	GA	GA	GA	X
12	0.911	0.705	0.901	0.868	0.608	0.856	0.727	0.859	LN	LN	LN	X
13	1.202	1.467	1.829	1.022	1.142	1.273	1.280	1.410	GA	GA	GA	X
14	0.891	0.937	1.000	0.858	0.792	0.902	0.871	0.914	LN	LN	LN	X,Y(20)
15	0.993	0.889	1.103	0.900	0.675	0.944	0.920	0.979	LN	LN	LN	X
16	2.584	1.372	7.480	1.397	1.162	2.390	1.389	2.183	(GA)	(GA)	GA	Z
17	1.149	1.014	1.244	0.957	0.762	0.966	0.963	1.025	GA	GU	GA,GU	X
18	0.849	0.530	0.892	0.811	0.442	0.822	0.590	0.785	(LN)	(LN)	LN	Y (20)
19	1.012	1.064	1.021	0.878	0.770	0.868	0.931	0.875	LN	LN	LN	X
20	0.868	0.907	1.049	0.808	0.687	0.897	0.785	0.854	LN	LN	LN	X,Y(23)
21	1.031	1.110	1.183	0.896	0.793	0.952	0.832	0.919	LN	GA	LN,GA	X
22	1.214	1.826	1.832	0.969	1.075	1.138	1.183	1.146	GA	GA	GA	X
23	1.007	0.791	1.087	0.875	0.593	0.983	0.670	0.947	LN	LN	LN	X
24	1.010	1.063	1.308	0.883	0.721	0.974	0.814	0.944	LN	GA	LN,GA	X
25	1.242	1.370	1.535	0.955	0.943	1.049	0.998	1.048	GA	GA,GU	GA,GU	X
26	1.340	2.300	2.204	1.016	1.296	1.255	1.279	1.217	GA	(GU)	GA	X
27	1.079	1.239	1.162	0.882	0.779	0.999	0.842	0.879	LN	GA	LN,GA	X
28	0.922	0.531	0.980	0.795	0.437	0.815	0.538	0.788	LN	LN	LN	X

Table 6.2. -Continued

STN NO	LOGNORMAL PDF			GAMMA PDF			GUMBEL PDF		PROBABLE BEST FIT BY			REMARKS
	ML	LS	MCS	ML	LS	MCS	LS	MCS	ML	WLS	ML&WLS	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
29	0.765	0.534	0.764	0.740	0.429	0.728	0.469	0.630	(LN)	(LN)	LN	Y (30)
30	0.990	0.933	1.268	0.852	0.621	0.928	0.664	0.850	LN	LN	LN	X
31	1.466	3.133	2.499	1.026	1.341	1.237	1.307	1.211	GA	(GU)	GA	X
32	0.897	0.633	0.976	0.810	0.473	0.818	0.462	0.723	LN	LN	LN	X
33	0.942	0.942	1.165	0.827	0.620	0.882	0.636	0.805	LN	LN	LN	X
34	1.453	2.943	3.028	1.014	1.265	1.378	1.173	1.297	GA	(GU)	GA	X
35	1.235	0.858	1.518	0.906	0.649	0.981	0.725	0.928	GA	LN	LN,GA	X
36	0.767	0.583	0.851	0.711	0.498	0.767	0.544	0.684	(LN)	(LN)	LN	Y (41)
37	1.046	1.450	1.259	0.850	0.822	0.888	0.833	0.851	LN	GA	LN,GA	X
38	1.803	2.388	3.675	1.066	1.235	1.332	1.102	1.243	GA	GU	GA,GU	X
39	1.407	1.844	2.092	0.961	0.972	1.090	0.940	1.048	GA	GA,GU	GA,GU	X
40	1.200	1.358	1.468	0.881	0.777	0.950	0.792	0.877	GA	GA	GA	X
41	0.917	0.468	0.956	0.786	0.345	0.795	0.359	0.691	LN	LN	LN	X
42	0.854	0.752	0.947	0.745	0.513	0.773	0.587	0.706	LN	LN	LN	X,Y(34)
43	1.378	2.038	1.969	0.945	0.944	1.040	0.910	0.997	GA	GA,GU	GA,GU	X
44	0.729	0.586	0.817	0.675	0.407	0.714	0.505	0.621	(LN)	(LN)	LN	Y (31)
45	1.023	0.642	1.322	0.806	0.441	0.958	0.536	0.745	LN	LN	LN	X
46	0.644	0.403	0.663	0.635	0.300	0.650	0.342	0.525	(LN)	(LN)	LN	Y (45)
47	1.746	2.401	3.262	1.033	1.006	1.301	0.859	1.224	GA	GA,GU	GA,GU	X
48	1.203	1.737	1.699	0.883	0.827	0.984	0.753	0.957	GA	GA	GA	X
49	1.113	1.365	1.165	0.821	0.710	0.843	0.732	0.783	LN	(GA)	LN,GA	X
50	1.413	1.041*	1.773	0.901	0.618	0.942	0.567	0.885	GA	GA	GA	X
51	0.752	0.556	0.816	0.674	0.350	0.707	0.409	0.601	(LN)	(LN)	LN	Y (38)
52	0.457	0.124	0.470	0.521	0.114	0.537	0.136	0.355	(GA)	(GA)	GA	Y (43)
53	0.753	0.664	0.900	0.674	0.425	0.727	0.419	0.651	(LN)	LN	LN	X,Y(46)
54	0.235	0.100	0.291	0.374	11	0.403	0.116	0.234	(GA)	(GA)	GA	Y (35)
55	2.332	6.506	7.647	1.065	1.364	1.486	0.975	1.266	GA	GU	GA,GU	X
56	0.689	0.344	0.834	0.612	0.250	0.673	0.307	0.546	(LN)	(LN)	LN	Y (29)

Table 6.2. -Continued

STN	LOGNORMAL PDF			GAMMA PDF			GUMBEL PDF		PROBABLE BEST FIT BY			
NO	ML	LS	MCS	ML	LS	MCS	LS	MCS	ML	WLS	ML&WLS	REMARKS
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
57	1.459	1.260	1.622	0.879	0.617	0.899	0.498	0.905	(GA)	GU	GA,GU	X
58	0.912	1.098	1.266	0.725	0.517	0.814	0.484	0.774	LN	LN	LN	X
59	1.472	1.065*	1.310	0.780	0.511	0.786	0.456	0.756	(GA)	(GA)	GA	Z (38)
60	1.048	0.678	1.159	0.745	0.324	0.777	0.310	0.772	LN	LN	LN	X
61	0.569	0.153	0.712	0.559	0.149	0.619	0.210	0.478	(LN)	(LN)	LN	Y (24)
62	0.794	0.320	0.745	0.636	0.214	0.633	0.234	0.544	(LN)	(LN)	LN	Y (36)
63	0.440	0.065	0.521	0.499	0.060	0.540	0.084	0.376	(GA)	(GA)	--	Y (21)
64	0.756	0.518	0.941	0.576	0.272	0.655	0.255	0.613	(LN)	LN	LN	X,Y(39)
65	0.790	0.352	0.755	0.579	0.233	0.573	0.233	0.460	(LN)	(LN)	LN	Y (35)
66	2.730	10.648	12.431	99	99	99	00	00	--	--	--	Z (25)
67	1.610	0.567	2.556	99	99	99	00	00	--	LN	LN	Z (30)

00 -SOLUTION NUMERICALLY UNSTABLE
11 -NO CONVERGENCE AFTER 50 ITERATIONS
99 -GA PDF NOT FIT
X -VARIANCE RATIO CLOSE TO UNITY(DEVIATION <.15 FROM UNITY) FOR ATLEAST ONE OF THE PDF'S
Y -VARIANCE RATIO <0.9 FOR LN AND GA BY ML, AND IN MOST CASES FOR THE THREE PDF'S BY WLS
Z -HARD TO FIT CASES
(GA) -IN COLUMNS (10) AND (11), FOR THE PDF IN THE PARENTHESES THE VARIANCE RATIO DIFFERS BY >0.15 FROM UNITY
LN -IN COLUMN (12), THE PDF WITH AN UNDERLINE INDICATES THE BEST PDF
(27) -IN COLUMN (13), THE NUMBER IN THE PARENTHESES INDICATES THE PERCENT VARIANCE CONTRIBUTED TO S BY K
* -THESE RATIOS WERE CONSIDERED TO BE LOW DUE TO BIAS IN LS METHOD(ϕ =0.0)

chosen does not make much difference.

The 50 "Easy-to-Fit" samples from Table 6.2 were further grouped according to the sample variance and suitability of a particular PDF, and the results are presented in Table 6.3. The GU distribution was excluded from this tabulation because it was not found superior for any single sample. Table 6.3 shows that the lognormal or the gamma distributions provided a good fit for a majority (75%) of flood samples. The other samples do not fit the two distributions well possibly because of outliers (see the next section). Table 6.3 also shows that samples with higher variances ($S_k^2 > 0.6$) are harder to fit. This may be because most high variance samples contain outliers (see next section).

At low variances ($S_k^2 < 0.2$), the GA distribution appears to be preferred to LN. At $S_k^2 = 0.2$ to 0.4 the "LN-Best" and "GA-Best" samples were about evenly divided. At $S_k^2 > 0.6$ the GA is not preferred which may be expected because as σ_k^2 approaches unity the GA distribution tends to an exponential (see Figure 3.2) which is unlikely to fit annual flood series data.

The simulation experiments (Table 5.2) also show that $\sigma^2_{F,LN.GA/ML}$ is, on the average, less than S_k^2 and $\sigma^2_{F,GA.LN/ML}$ is, on the average, greater than S_k^2. Such occurrences were also postulated in Chapter IV. The results presented in Table 6.2 show that $\sigma^2_{F,(LN).GA/ML}$ is less than S_k^2 while $\sigma^2_{F,(GA).LN/ML}$ is greater than S_k^2.

The 100-year predictions given by the methods of MO and ML when the "LN Best" and "GA Best" samples of Table 6.2 are fit to both LN and GA are presented in Tables 6.4 and 6.5, respectively. These results

Table 6.3. Analysis of Easy - to - Fit Samples

S_K^2	No of Cases Examined	$\sigma_F^2 \simeq S_K^2$ for one of the PDF's		Best PDF		
		No of Cases	Percent	LN	GA	BOTH
0.2	18	16	89	5	8	3
0.2 to 0.4	26	23	88	13	10	0
0.41 to 0.6	14	9	64	4	5	0
0.61 to 1.0	7	2	29	2	0	0
1.0	2	0	0	0	0	0
Total	67	50	75	24	23	3

account for two facts that have been observed for several years. First, the predictions differ greatly when the same data sample is subject to frequency analysis by different PDF's. Tables 6.4 and 6.5 show that the large differences obtained in predictions when a data sample is fit to different PDF's are due to both the choice of PDF and the choice of the computational method (MO, ML, etc.). With respect to lognormal, hydrologists have invariably used the ML method when fitting a LN distribution because the computations involved are simple (same computations as moment fit of logs to a normal distribution, see Appendix B).

The results of the simulation experiments presented in Chapter V, (Study No. 1) show that the shape fitting methods would produce results (K's for various return periods) greatly different from the results based on the parent PDF when GA or GU samples at high variance ($S_k^2 > 0.2$) were fit to LN data (see Table 5.3). For real samples also, Table 6.5 shows that $K_{S100,(GA).LN/ML}$ values are about 16% to 55% higher than $K_{S100,(GA)/(ML\ or\ MO)}$ for the same sample, depending on the value of S_k^2. The large disparity in predictions results, as shown by Table 6.4, from fitting a data (GU or GA) sample to a wrong (LN) distribution by a shape fitting (ML) method. The disparities in 100-year predictions obtained when "LN-Best" samples were fit to GA by ML are somewhat milder in nature and ranged from 10 to 21% for samples in the variance range of 0.2 to 0.7 (see Table 6.5). For samples whose variance is less than 0.2 the magnitudes of the disparities may be expected to be less than those shown in Tables 6.4 and 6.5. Tables 6.4 and 6.5 also show that

Table 6.4. 100 - Year Predictions of 'GA - Best' Samples ($S_K^2 > 0.2$)

STN NO	S_K^2	K_{S100} - GA		K_{S100} - LN		$\frac{K_{S100,(GA)LN/ML}}{K_{S100,(GA)GA/ML}}$
		MO	ML	MO	ML	
22	0.2348	2.44	2.43	2.62	2.83	1.16
25	0.2519	2.51	2.48	2.70	2.94	1.19
26	0.2567	2.52	2.55	2.71	3.05	1.20
31	0.3053	2.70	2.73	2.90	3.43	1.26
34	0.3196	2.75	2.77	2.96	3.48	1.26
35	0.3329	2.80	2.69	3.01	3.30	1.23
38	0.3530	2.86	2.94	3.08	4.01	1.36
39	0.3633	2.87	2.85	3.12	3.63	1.27
40	0.3657	2.90	2.76	3.13	3.40	1.23
43	0.3850	2.97	2.90	3.20	3.69	1.27
47	0.4048	3.03	3.07	3.27	4.21	1.37
48	0.4116	3.05	2.90	3.29	3.58	1.23
50	0.4704	3.23	3.09	3.49	4.09	1.32
55	0.5462	3.46	3.55	3.74	5.50	1.55
57	0.5851	3.56	3.36	3.83	4.58	1.36

Table 6.5. 100 - Year Predictions of 'LN - Best' Samples ($S_K^2 > 0.2$)

STN NO	S_K^2	K_{S100} - LN		K_{S100} - GA		$\frac{K_{S100,(LN)GA/ML}}{K_{S100,(LN)LN/ML}}$
		MO	ML	MO	ML	
19	0.2120	2.51	2.53	2.35	2.25	0.89
20	0.2241	2.57	2.44	2.40	2.24	0.92
21	0.2248	2.58	2.61	2.40	2.33	0.89
23	0.2395	2.63	2.65	2.45	2.35	0.89
24	0.2448	2.66	2.68	2.48	2.38	0.89
27	0.2756	2.78	2.88	2.50	2.49	0.86
28	0.2775	2.78	2.71	2.50	2.40	0.89
30	0.2943	2.86	2.85	2.55	2.51	0.88
32	0.3055	2.90	2.78	2.70	2.50	0.90
33	0.3103	2.92	2.86	2.72	2.54	0.89
37	0.3487	3.08	3.13	2.85	2.69	0.86
41	0.3750	3.18	3.05	2.93	2.67	0.88
42	0.3837	3.20	3.01	2.97	2.64	0.88
45	0.4032	3.27	3.30	3.03	2.77	0.84
49	0.4315	3.37	3.54	3.11	2.87	0.81
53	0.5974	3.83	3.73	3.59	3.12	0.84
60	0.7043	4.18	4.27	3.88	3.39	0.79
64	0.9513	4.81	4.24	4.51	3.47	0.82

the disparities in 100-year predictions when the "best PDF" was not chosen are less when the method of moments is used. Thus the errors introduced in not choosing the "best PDF" are larger when computations are made using the so-called statistically superior ML method than when the computations are made by the method of moments. However, by introducing large errors into the solution ML, perhaps, cautions a hydrologist that he chose a wrong PDF!

In Tables 6.4 and 6.5, since the samples were arranged in ascending order of variance magnitude, one might expect the disparities in results to show similar order. However, results shown in column 7 of Tables 6.4 and 6.5 displays a trend with noise. One reason for this occurrence may be that the real samples, unlike simulated samples, are not samples from known populations.

Another consideration is that predictions given by MO and by ML sometimes differ greatly when the same PDF is used (see LN results of stations 38, 47 and 55, Table 6.5). In such instances, a statistician-hydrologist might react by disregarding the method of moments as inferior. Tables 6.4 and 6.5 suggest, instead, that one reason for disparities in MO and ML predictions could be that the procedure used tried to fit the data sample to a distribution of substantially different shape than the sample and that, in fact, in such cases, the MO fit yields better predictions.

As a next step in the analysis of the "Easy-to-Fit" samples, one each from Table 6.4 (GA best fit 43) and Table 6.5 (LN best fit 32), were plotted on log-probability paper (Figures 6.1 and 6.2) using for

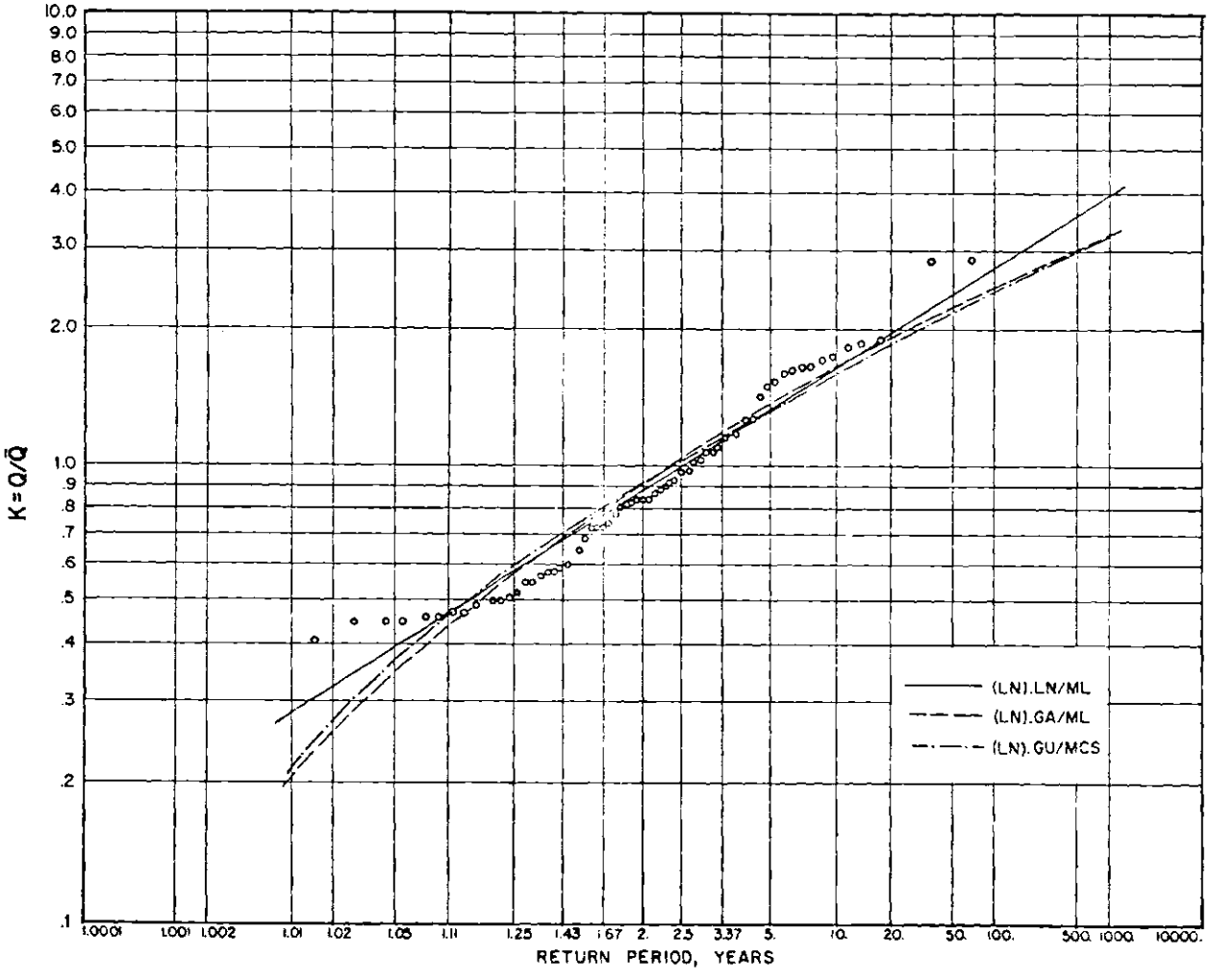


Figure 6.1 LN 'Best' Sample (Station No. 43: Chattahoochee River Near Norcross, Georgia)

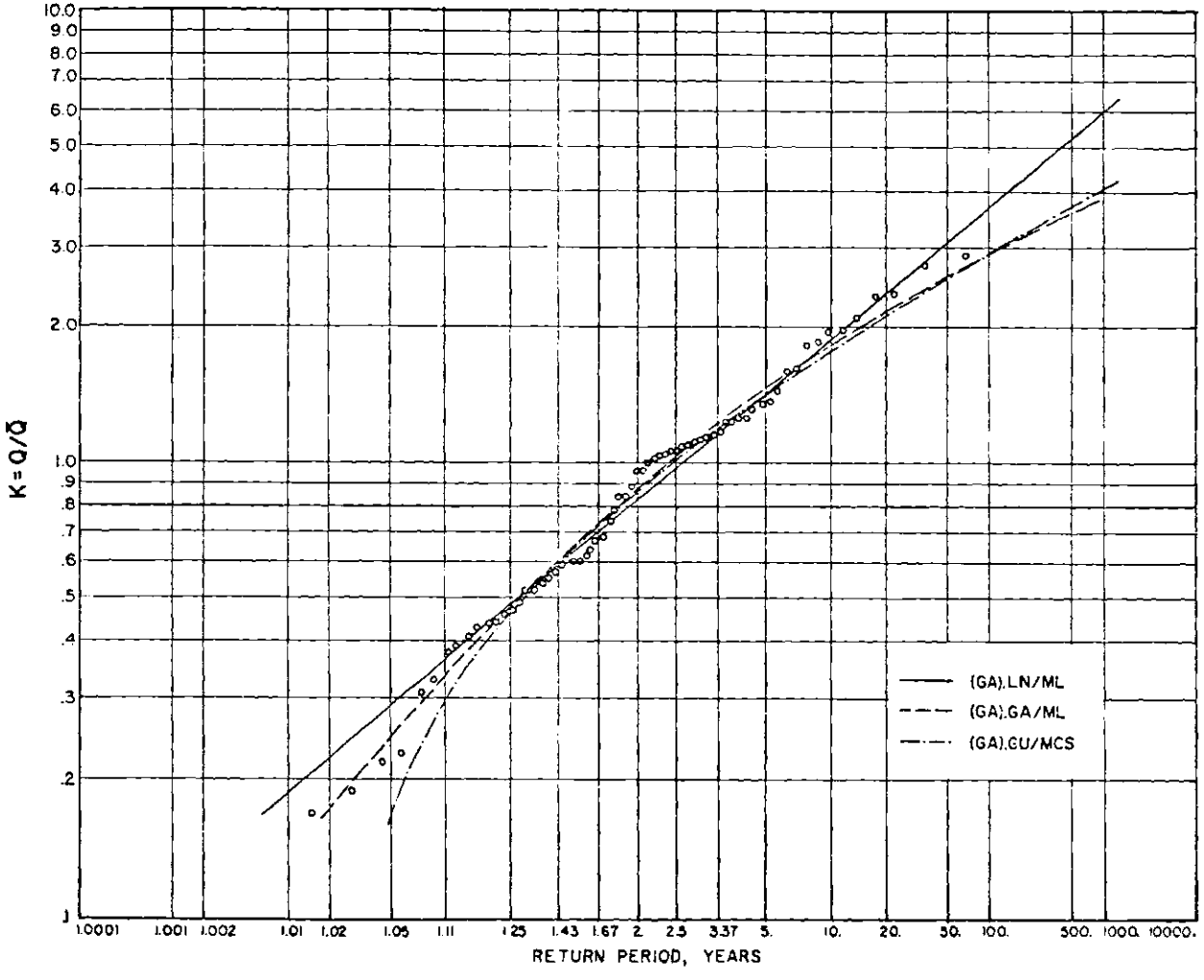


Figure 6.2 GA 'Best' Sample (Station No. 43: Trinity River at Riverside, Texas)

plotting positions the well known Weibull formula:

$$P(K \geq k) = \frac{m}{n+1} \tag{6.2}$$

in which m is the order of K (m = 1 corresponding to the largest K) and n is the number of years of data.

Figure 6.1 shows that the data for station 32 appear to be better fit by LN than by GA or GU. Similarly, Figure 6.2, on which the data of station 43 are plotted, shows that the data appear to be better fit by GA (or GU with extreme lower tail ignored, which were shown as the best PDF's by the 'Best Fit' criterion) than LN. In Figure 6.1 the data, on the whole, plot as a straight line whereas the GA and GU fits, though computed by shape fitting methods, have a convex curvature and are, thus, not suitable. In Figure 6.2 the data, in general, plot with a convex curvature whereas the LN-ML fit shown, though computed by a shape fitting method, follows a straight line and is, thus, not suitable. Figure 6.2 also shows that, since the lower tail of the GU distribution extends to $-\infty$ theoretically, it does not fit real data in the lower tail for higher variance samples. Figures 6.1 and 6.2 demonstrate that the PDF's selected by the "Best Fit" criterion coincide with the best visual fit among the PDF's examined when the data is plotted on probability paper.

Samples with Outliers

In the literature of hydrology, outliers are defined as data items "more far removed from the trend of the others," (Bulletin No. 13, Water Resources Council, 1966). Beard, in recent work on flow frequency

analysis (1974), defined outliers as "extreme values whose ratio to the next most extreme value in the same (positive or negative) direction is more extreme than the ratio of that next most extreme value to the eighth most extreme value." Outliers, being "more far" removed from the other data points, contribute a large part of the square of the departure from the mean used to compute the sample variance, S_k^2. Thus, outliers account for an unusually large fraction of sample variance.

Outliers in Normally Distributed Samples. Nair and Chauvenet (Kennedy and Neville, 1976) have each devised a criterion to identify outliers in normal populations.

According to Nair, the maximum deviation from the sample mean $\overline{X}$ that can be expected for single values in samples of size n is related to the estimated variance of the population; the estimate must be based on a larger sample than the one containing the outlier. However, Nair's method is limited to samples of sizes 3 to 9 only. Nair's method shows that for a sample of size 9 drawn from a population of size ∞, at 5% significance level, an outlier will contribute at least 72% variance . Since Nair's method is restricted to very small samples it is not useful for application to hydrologic samples of the sizes used in this study.

According to Chauvenet, an observation in a sample of size n is regarded as an outlier if it has a deviation from the mean greater than a 1/(2n) probability. The probability is calculated on the assumption of a normal distribution and sample variance. For example, if n = 10, then 1/(2n) = 0.05, which is the probability of a normal deviate of 1.96σ. Thus an observation which deviates from the mean by at least

1.96S is regarded as an outlier in ten years of record. The following table gives, for different sample sizes, the minimum percent variance to be contributed by $K_{(m=1)}$ to qualify the value as an outlier (Min % $V_{O/L}$) by Chauvenet's criterion.

n	$Z = \frac{K_{m=1} - \bar{K}}{S_k}$	Min % $V_{O/L}$
10	1.96	43
20	2.24	26
30	2.39	20
40	2.50	16
50	2.58	14
60	2.64	12
80	2.74	10
100	2.81	9

(In the above table, the equation to compute 'min % $V_{O/L}$' may be derived as follows:

By analogy with Equation 6.1,

$$\text{min \% } V_{O/L} = \left[\frac{(K_{m=1} - \bar{K})^2}{n-1} / S_k^2\right] \times 100 \qquad (6.3)$$

Substituting $Z = \frac{K_{m=1} - \bar{K}}{S_k}$ in 6.3 yields

$$\text{min \% } V_{O/L} = \frac{Z^2}{n-1} \times 100 \qquad (6.4)$$

The above table may be extended to any values of n). Chauvenet suggests to reject an outlier (based on the above table) from the sample. However, such a procedure could easily lead to successive rejection of extreme observations -- a procedure that must never be used. Rejecting an observation may also mean throwing away vital information about the sample. Since most hydrologic samples are from skewed populations, however, further analysis is needed for most hydrologic applications.

Outliers in Skewed Populations. Chauvenet's criterion shows that an observation in the upper tail of a normal distribution will be considered as an outlier if the exceedence probability for the observation is less than or equal to 1/(4n) and the distribution has a variance equal to the sample variance. Thus, observations estimated to have a return period exceeding 4n years by the fit to a (normal) distribution are outliers. This arbitrary definition for outliers can be extended to the LN and GA distributions. Using this extension of his method, minimum % $V_{O/L}$'s were computed for σ_k^2 = 0.1, 0.2, 0.3, 0.5, 0.7, and 1.00. The results are shown graphically by Figure 6.3.

Figure 6.3 shows that the value of min % $V_{O/L}$ depends on the sample size and skewness of a distribution. At a given σ_k^2, LN has a larger skew than GA and skewness of both LN and GA increases as σ_k^2 increases (see Table 3.2). Figure 6.3 shows similar variations in min % $V_{O/L}$ also. Figure 6.3 also shows that an observation which may be regarded as an outlier by GA need not be an outlier by LN.

By the criterion developed in Figure 6.3, 13% of GA (synthetic)

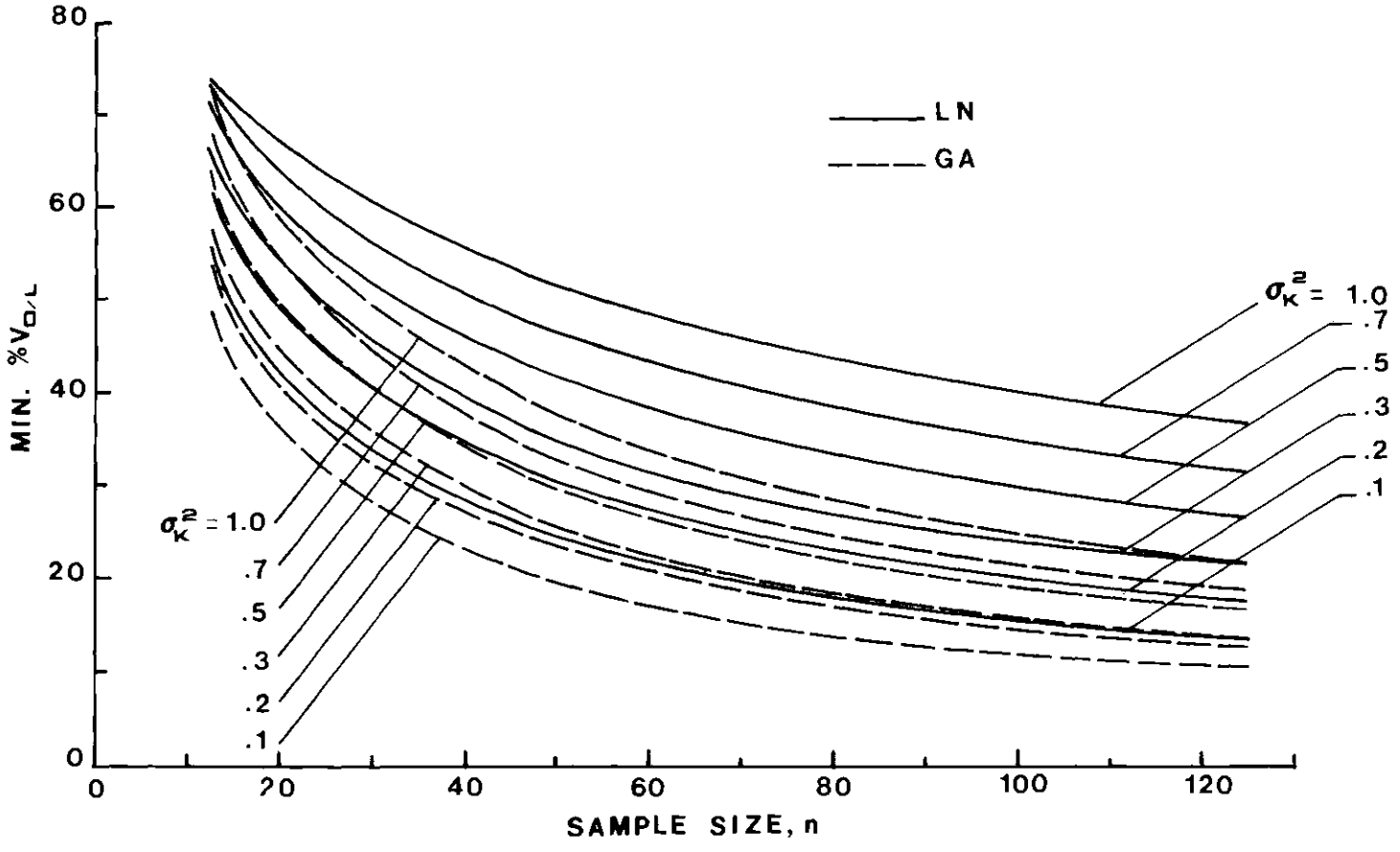


Figure 6.3 Minimum percent variance due to the largest observation (Km=1) needed to qualify the value as outlier

samples and 19% of LN samples used in this study (see Chapter V) were found to contain outliers. As one might expect, for most of these samples S_k^2 was found to be greater than σ_k^2.

When such data points occur in a sample, and the sample is fit to a PDF by one of the shape fitting methods, σ_F^2 would be, in general, less than the sample variance because the shape fitting methods would give reduced weight to the data point "far removed" from the others and fit the rest of the data points, which together may have a variance much less than the total sample variance. While the largest value in a sample may be several times larger than the mean (for the 67 real samples of Table 6.1 the largest values were in a range of 1.6 to 7.34 times the sample mean) the smallest value in a sample cannot be less than zero or one unit "removed away" from the sample mean. (For the 67 real samples of Table 6.1 the lowest data points ranged from 0.03 to 0.61 of the sample mean). Also, while the largest data point may be removed from the next largest (or from the bulk of the data points) by several times the sample mean the lowest data point is removed from the next smallest (or bulk of the data points) only by a small fraction of the sample mean. Thus a single low value will usually not substantially contribute to the variance of the sample and thus its presence in a data sample will not strongly affect the fit by a shape fitting method. Note that the definition of S_k^2 is confined to natural data only but not to the logarithms of data. However, if several low values occur in a data sample the variance of the LN/GA PDF's fitted by shape fitting methods may be higher than the sample variance because these distributions

have a thicker lower tail only at higher variance (see Chapter III).

Samples with higher outliers: The percent contribution to the sample variance by the highest values ($K_{m=1}$) in the 67 real samples ranged from 7 to 85 (see Table 6.1). Table 6.6 shows that in 22 of the 67 samples the highest flow contributed 25% or more to the sample variance. High flows were found to be outliers for GA or for both GA and LN by Figure 6.3 (see Table 6.1). Table 6.6 also shows that 'outliers' are more prevalent in samples with large variance ($S_k^2 > 0.4$), where they occur in about 69% of the samples. Generally, the incidence of large variance in a (hydrologic) sample may mainly be due to the presence of an outlier.

Application of 'Best Fit' Criterion to Samples with Outliers. The 'Best Fit' criterion proposed in this study may not be suitable for samples with outliers because the variance ratio σ_F^2/S_k^2 will almost invariably be less than unity for any hypothesized PDF. While several techniques for handling outliers have been suggested in the past (Bulletin Nos. 13 and 15, see Bibliography) one technique used by hydrologists appears to be "to keep the value as is" (Beard, 1974). In such cases, for a given hypothesized PDF, the method of moments "retains" the full variance effect of the higher outliers while the shape fitting methods reduce or "correct" for the effect. One way to determine the PDF of best fit for samples with outliers appears to be to choose that PDF for which σ_F^2/S_k^2 is nearest unity, where σ_F^2 is evaluated by a shape fitting method (ML/LS/MCS), i.e., the 'Best Fit' criterion may still be applied to the samples with outliers. However,

Table 6.6. Percent Variance due to K_{max} - Flood Data

% Variance due to K_{max}	Number of Samples	
	$S_K^2 \leq 0.4$	$S_k^2 > 0.4$
8	2	0
9 - 16	21	3
17 - 24	13	6
25 - 32	6	5
33 - 40	1	5
41 - 48	1	3
49 - 56	0	0
57 - 64	0	0
65 - 72	0	0
73 - 80	0	0
81 - 88	0	1
Total	44	23

it should be realized that whether the computations are made by the method of moments or by shape fitting methods, the return period assigned to the outlier ($K_{m=1}$) on the basis of the selected PDF may be several times the period of record. It may also be noted that the predictions based on a shape fitting method will always be less than those based on the method of moments for samples with higher outliers (since $\sigma_F^2 < \sigma_K^2$ for shape fitting methods).

Since $\sigma^2_{F,ML/LS/MCS}$ is substantially less than S_k^2 for samples with outliers, examination of samples with σ_F^2/S_k^2 less than 0.9 is worthwhile. In Table 6.2, the samples for which the variance ratio, ML σ_F^2/S_k^2, is less than 0.9 are designated by the letter Y in the remarks column together with the percent contribution to the sample variance by the highest flows in the respective samples. For these samples, Table 6.7 shows that the variance contributed by the highest flood ($K_{m=1}$) is 20% or greater and the total variance contributed by the highest two or three data items ($K_{m=1}$, $K_{m=2}$, $K_{m=3}$) which were close to each other in value varied from 31% to 85%. Table 6.7 also shows that in samples with low S_K^2, the outliers do not affect $\sigma^2_{F,ML}$ greatly (see the values of $\sigma^2_{F,ML}/S_K^2$ for stations 5, 14, 18 and 20). As the variance of the samples increases the value of $\sigma^2_{F,ML}/S_K^2$ decreases and, in general, depends on the variance contributed by the outlier(s).

When $K_{m=1}$ and $K_{m=2}$ are closer in magnitude, Beard excludes such high flows including the highest from the definition of outliers (see Beard's definition for outliers stated earlier). Table 6.7 shows that even when two or three very high flows of similar magnitude occur close to each other and together contribute substantially to S_K^2 the effect of

Table 6.7. Results of Samples with (Higher) Outliers

STN NO (1)	n (2)	S (3)	% VARIANCE DUE TO $K_{m=1}$ (4)	% VARIANCE DUE TO $K_{m=2}$ (5)	$\sigma^2_{F,ML}/S^2_K$ LN (6)	$\sigma^2_{F,ML}/S^2_K$ GA (7)	K_{S100} -LN MO (8)	K_{S100} -LN ML (9)	K_{S100} -GA MO (10)	K_{S100} -GA ML (11)	PROBABLE BEST PDF (12)
5*	57	0.0929	27	13	0.875	0.894	1.92	1.84	1.84	1.79	GA,LN
14	56	0.1512	20	18	0.891	0.858	2.22	2.14	2.10	2.02	LN
18	60	0.1880	20	11	0.849	0.811	2.41	2.27	2.27	2.12	LN
20@	61	0.2241	23	21	0.858	0.808	2.57	2.44	2.40	2.24	LN
29@	66	0.2781	30	13	0.765	0.740	2.80	2.52	2.61	2.34	LN
36*	62	0.3359	41	10	0.767	0.711	3.02	2.71	2.80	2.47	LN
42*	72	0.3837	34	10	0.854	0.746	3.20	3.01	2.97	2.64	LN
44*	72	0.4007	31	21	0.729	0.675	3.25	2.85	3.02	2.58	LN
45*	61	0.4044	45	12	0.644	0.635	3.27	2.72	3.03	2.54	LN
51*	63	0.4851	33	12	0.752	0.674	3.54	3.13	3.28	2.73	LN
52*	56	0.5009	43	20	0.457	0.521	3.59	2.58	3.32	2.55	GA
53*	54	0.5075	46	6	0.753	0.671	3.63	3.19	3.34	2.83	LN
54*	58	0.5182	35	4	0.285	0.374	3.65	2.20	3.35	2.30	GA
56@	66	0.5548	29	26(14)	0.689	0.612	3.75	3.19	3.48	2.92	LN
61	52	0.7063	24	23(20)	0.569	0.559	4.18	3.25	3.88	3.00	LN
62@	52	0.8775	36	14	0.794	0.636	4.38	3.94	4.06	3.30	LN
63	53	0.7681	21	21(18)	0.440	0.499	4.60	3.17	4.28	3.12	GA
64@	62	0.9613	39	23	0.756	0.576	4.81	4.24	4.51	3.47	LN
65@	56	0.9738	35	35	0.790	0.579	4.84	4.35	4.54	3.50	LN

FIGURES IN PARENTHESES IN COLUMN (5) INDICATE THE % VARIANCE DUE TO $K_{m=3}$

* FOR THESE SAMPLES $K_{m=1}$ IS AN OUTLIER FOR BOTH LN AND GA BY FIGURE 6.3

@ FOR THESE SAMPLES $K_{m=1}$ IS AN OUTLIER FOR GA, BUT NOT FOR LN BY FIGURE 6.3

the presence of such multiple high flows in the sample is, in general, the same as the presence of a single outlier as far as the fit by ML is concerned. However, the effect of a single outlier having a large variance contribution appears to be more pronounced than the effect of a group of high flows having a total contribution similar to the individual contribution (see the magnitude of σ_F^2/S_K^2 of station 54 compared to that of stations 56 and 65). Table 6.7 shows that the LN PDF is more generally selected by the variance ratio 'Best Fit' criterion when fitting data with high outliers.

By the criterion of Figure 6.3, $K_{m=1}$ was found to be an outlier for all samples except stations 14, 18, 61 and 63 of Table 6.7 by GA distributions. By LN distributions for about one-half of the samples $K_{m=1}$ was found to be an outlier. These results show that, in general, when an outlier defined by Figure 6.3 occurs in the sample, the variance ratio σ_F^2/S_K^2 is much less than (< 0.9) unity.

Table 6.7 also compares the ML and MO predictions for the 100-year return period for samples with outliers. The differences in the sample variance and $\sigma_{F,ML}^2$ are clearly reflected in predictions by the two computational methods. At the outset, the MO predictions may appear more "commensurate" with the actual highest flow observed and, perhaps, they may be used where a conservative estimate is desired. On the other hand, if it is recognized that the observed highest flow is an exceptionally rare event and inclusion of its full effect might give highly conservative estimates, the results given by ML method may be used. In any event, the hydrologists should note that the computational methods will make a difference when dealing with samples with outliers.

The least squares method with $\phi = 0.0$ generally gives a very poor fit (i.e., $\sigma_F^2 << S_k^2$) when higher outliers are present in the sample and use of a weight exponent (ϕ) close to 1.0 (MCS method) would invariably be required to obtain better fit (See the results of stations marked Y in Table 6.2.).

Samples with Extreme Lower Values. Hydrologists regard both higher and lower extreme values as outliers. The term "outlier" appears to be somewhat a misnomer when applied to the lower extreme values of a data sample. They are not "more far removed" from the others. Their contribution to the sample variance is not significant. Unlike the higher outliers, they may affect the fit (by shape fitting methods) only when they occur in multiple numbers. It appears more appropriate to confine the term "outlier" solely to data items identified by a large contribution (25% or higher) to the sample variance. To examine whether the lower extreme values have any general influence on the variance ratio $\sigma^2_{F,ML/LS/MCS}/S_K^2$ of PDF's fitted, Beard's definition of (lower) outliers was first applied to the 67 real samples and the samples with lower outliers were separated. By Beard's definition, the lowest flows of stations 12, 16, 21, 32, 34, 51, and 56 are found to be outliers of magnitudes .33, .06, .30, .41, .15, .24, and .13, respectively. The variance ratios presented in Table 6.2 for the above stations show that, with the exception of station 16, these outliers do not affect σ_F^2/S_K^2 in a general way for the PDF's applied. In case of stations 12, 21, 32, and 34 the "best fit" criterion could be applied and the "best" PDF

determined. Stations 51 and 56 were identified as stations with (higher) "outliers" (see previous section). Station 16 is discussed in detail in the next section. The magnitudes of the outliers themselves indicate that they cannot be considered as far removed from the rest of the data though they fit into the empirical definition proposed by Beard.

The samples for stations 11, 38 and 55 (Table 6.2) though indicated as "Easy-to-fit samples" (the gamma distribution has been found to fit them "best") are the samples with multiple lower extreme values. For these samples, the variance ratios by all the shape fitting methods are, in general, larger than unity for the three hypothesized PDF's (see Table 6.2). Since the lower extreme values do not distinguish themselves with a reference to their contribution to the sample variance, one way to distinguish the lower extreme values as anomalous appears to be with respect to their position relative to certain lower percentiles like $K_{.01}$, $K_{.05}$, $K_{.10}$, etc., of the hypothesized PDF. For the samples from station numbers 11, 38 and 55, the flow terms occurring at the lower tails of LN and GA distributions with $\sigma_F^2=S_K^2$ (MO fit) were analyzed with reference to their occurrence relative to the 1, 5 and 10 percent percentiles ($K_{.01}$, $K_{.05}$, $K_{.10}$) and the results are presented in Table 6.8. Table 6.8 shows that the observed occurrences of low flows were several times the expected occurrences with respect to LN distribution at one and five percent percentiles, $K_{.01}$ and $K_{.05}$, for the three samples. Such an occurrence in low flows renders the lower tail of the sample distribution much thicker than that of the proposed PDF (LN) with

Table 6.8. Analysis of Data in Extreme Lower Tail of PDF's

Stn No	S_K^2	Sample Size	No. of Flows Smaller Than or Equal to the Indicated Percentile -- Expected -- $K_{.01}$	$K_{.05}$	$K_{.10}$	-- Observed - LN* -- $K_{.01}$	$K_{.05}$	$K_{.10}$	-- Observed - GA* -- $K_{.01}$	$K_{.05}$	$K_{.10}$	$(\sigma_F^2/S_K^2)_{ML}$ LN	GA
11	0.1358	77	1	4	8	2 (0.41)	8 (0.52)	12 (0.59)	2 (0.34)	6 (0.48)	11 (0.57)	1.286	1.064
38	0.3530	66	1	3	7	6 (0.24)	7 (0.35)	10 (0.43)	1 (0.13)	7 (0.26)	7 (0.35)	1.803	1.066
55	0.5462	64	1	3	6	5 (0.17)	12 (0.27)	13 (0.35)	0 (0.06)	5 (0.16)	10 (0.24)	2.332	1.065

* LN and GA distributions based on variance equal to the sample variance

(0.41) The figures in parentheses indicate magnitudes of the percentiles

$\sigma_F^2 = S_K^2$ and the computational methods like ML and LS, in an attempt to fit the shape, result in a fit with variance much larger than the sample variance in case of lognormal distribution. The GA distribution which has, in general, a thicker lower tail appears to be more suitable for application for samples with multiple extreme lower flows. For the above three samples, with respect to GA, the expected and the observed low flows are about the same at the percentile $K_{.01}$, and differences between the observed and the expected occurrences are less compared to LN. The LS and MCS fits for GA distribution showed a much larger variance than ML fit for the above three samples (see Table 6.2). From the above occurrences it may be concluded that if a sample contains lower extreme values in groups $\sigma^2_{F,ML/LS/MCS}/S_K^2$ will be greater than unity if LN or GA PDF's are applied. The effect on predictions will be to increase the predictions given by the shape fitting methods.

Table 6.9 presents the MO and ML 100-year sample predictions (K_{S100}) when the LN and the GA distributions are applied to the three samples with (multiple) lower values. Table 6.9 shows that the ML predictions by LN are highly exaggerated compared to those given by the method of moments. While the GA distribution may be regarded as the most applicable PDF for samples with low outliers, the errors introduced in not choosing the "best fit" will be minimal when computations are made by the method of moments when LN is applied (see Table 6.9). In general, when LN is applied to samples with groups of extreme lower values, the shape fitting methods result in highly exaggerated predictions.

The discussion presented in this section is generally applicable

Table 6.9. Samples with Data in Extreme Lower Tail of PDF's - Predictions

Stn No	S_K^2	$\sigma^2_{F,ML}/S_K^2$		K_{S100} - LN		K_{S100} - GA		Probable Best PDF
		LN	GA	MO	ML	MO	ML	
11	0.1358	1.286	1.064	2.15	2.35	2.04	2.09	GA
38	0.3530	1.803	1.066	3.08	4.01	2.86	2.94	GA
55	0.5462	2.332	1.065	3.74	5.50	3.46	3.55	GA

to samples with a positive skew and may not be strictly applicable to samples with a negative skew. Hydrologic samples are generally positively skewed samples.

Hard-to-Fit Samples

The flood samples which do not readily fit into the definition of either "easy-to-fit" or "samples with outliers" are termed "hard-to-fit samples." The four of the 67 samples (marked Z in Table 6.2) so classified are discussed below individually.

Blue River at Dillon Colorado (Station 16). The histogram of this sample (Figure 6.4) shows a negative skew (g = -.046) while each of the three PDF's used in this work has a positive skew (See Chapter III.) and thus obviously cannot fit the sample. No detailed investigation as to what would happen if a positively skewed distribution is fit to a negatively skewed sample by shape fitting methods was made in this study; but for this station, μ_F and σ_F^2/S_K^2 when LN, GA, and GU were applied are

Blue River at Dillon, Colorado

	μ_F*			σ_F^2/S_K^2		
PDF	ML	LS	MC	ML	LS	MCS
LN	1.047	1.181	1.410	2.584	1.372	7.480
GA	1.000	1.126	1.163	1.397	1.162	2.390
GU	-	1.156	1.151	-	1.389	2.183

* $\bar{X} = 1$

Thus, at least in this case, both the mean and the variance of

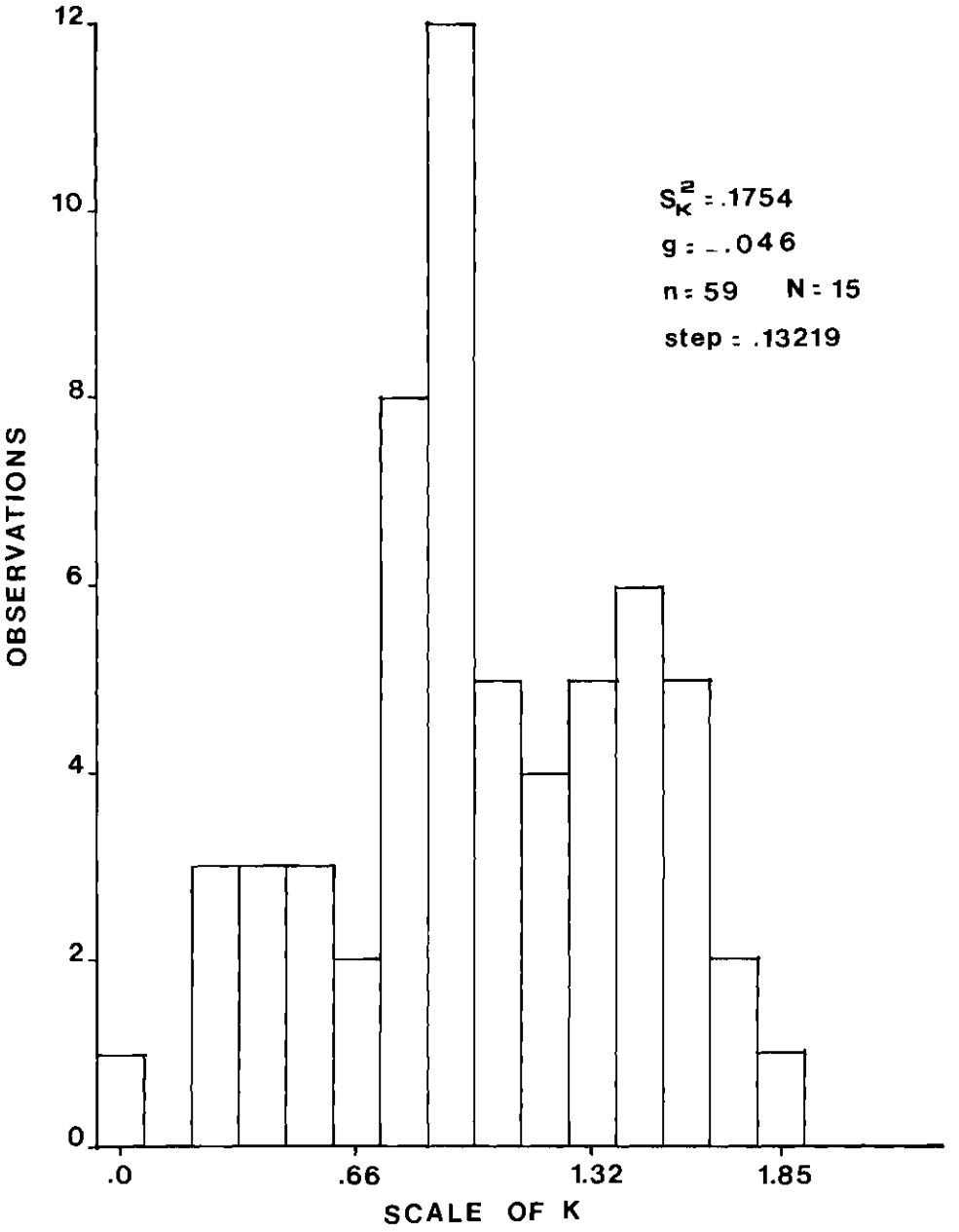


Figure 6.4 Histogram of Annual Peak Flow, Blue River at Dillon, Co.

the fitted PDF were much higher than the sample values when a positively skewed PDF was applied to negatively skewed data by a shape fitting method.

Emigration Creek Near Salt Lake City, Utah (Station 59). The histogram is presented in Figure 6.5. For this sample, Table 6.2 shows that even though the sample contained a high outlier that contributed 38% of the sample variance, σ_F^2/S_K^2 for LN PDF was 1.47 (i.e., larger than 1.0). One reason could be that the sample contains many very low data items whose influence dominate the higher outliers in the case of the LN by ML/LS/MCS methods. The sample (n = 57) contains two values of magnitude 0.06. (The percentile $K_{.01}$ for LN, with $\sigma_K^2 = 0.7$, had a magnitude of 0.14). For this sample the GA distribution may be regarded as the best fit (The percentiles $K_{.01}$ and $K_{.05}$ had magnitudes .03 and 0.11 respectively at $\sigma_K^2 = 0.7$, and thus the two lowest values are not unusually low for the GA distribution) and $\sigma_{F,ML}^2/S_K^2$ was substantially less than 1.0 due to the influence of the (high) outlier. The 100-year predictions (K_{S100}) of MO and ML methods by LN and GA distributions for this sample are as follows:

Emigration Creek Near Salt Lake City, Utah

PDF	K_{S100} MO	K_{S100} ML
LN	4.17	4.99
GA	3.87	3.45

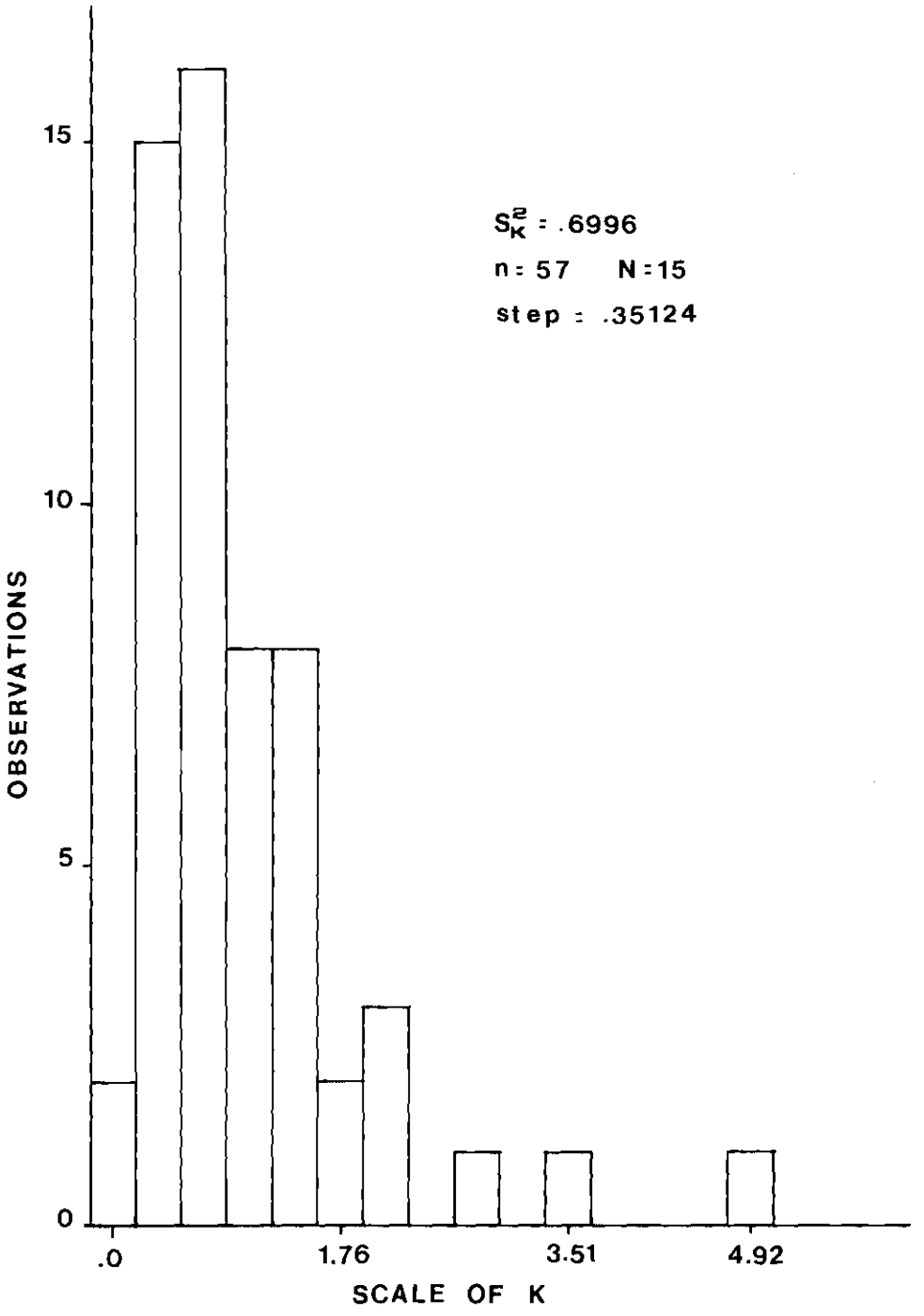


Figure 6.5 Histogram of Annual Peak Flows, Emigration Creek at Salt Lake City, Utah

In this example, when both higher outliers and lower values are present, the influence of lower outliers is greater when LN distribution is fit by shape fitting methods. It was not investigated in this study whether such an occurrence is general.

Tule River Near Porterville and Arroyo Seco Near Pasadena, California (Stations 66 and 67, Respectively). These two samples (Figure 6.6) have a sample variance larger than unity; thus, the gamma and the Gumbel distributions are not suitable for application. These two samples each have two high outliers which together contribute 47% and 59% to the variance at stations 66 and 67, respectively. However, Table 6.2 shows that the values of $\sigma^2_{F,ML}/S^2_K$ are of the order of 2.7 and 1.6 for these two samples, a phenomenon which should not have occurred if only higher outliers were present in the samples. The samples also contained the low values:

		No. of Flows Below or Equal to the Percentile			
Stn.	Sample Size	Expected		Observed - LN*	
		$K_{.01}$	$K_{.05}$	$K_{.01}$	$K_{.05}$
Tule River	64	1	3	4 (.07)	11 (.14)
Arroyo Seco	55	1	3	3 (.04)	6 (.09)

* LN based on variance equal to the sample variance. Values in parentheses indicate the magnitude of percentile.

The influence of the lower outliers is dominant on the LN fit by the shape fitting methods (See discussion on Station 59.) and thus $\sigma^2_{F,ML}/S^2_k$ is found to be much larger than 1.0 for the two samples. When samples with such thicker lower tails are encountered, search should be

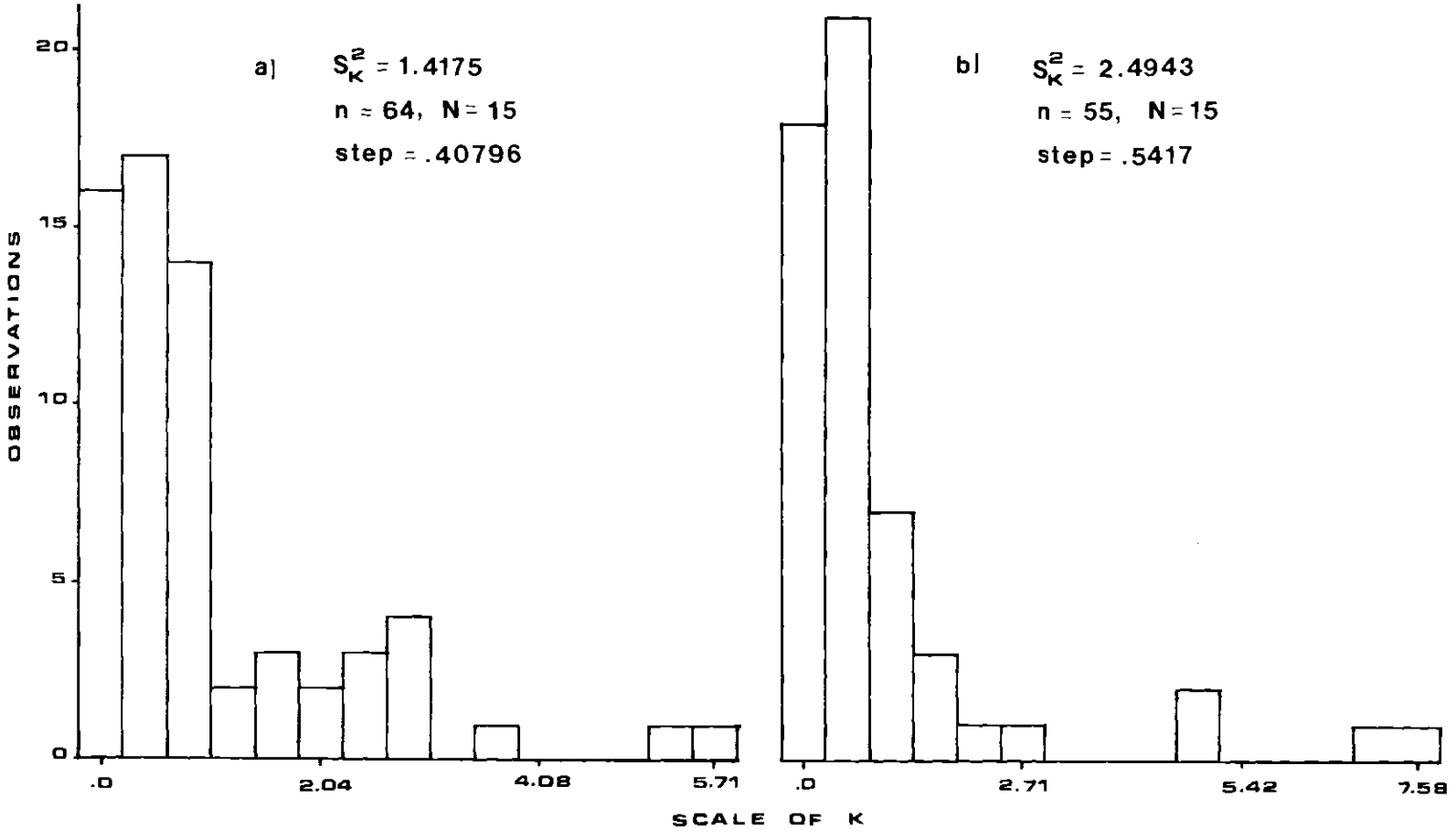


Figure 6.6 Histogram of Annual Peak Flows
(a) Tule River near Portersville, Ca.
(b) Arroyo Seco near Pasadena, Ca.

made for PDF's which have thicker lower tails and those PDF's should be used in frequency analysis. If LN is applied to such samples MO method may be preferred to ML because of the tendency of the latter to result in a larger σ_F^2 than S_k^2. The 100-year predicted flows for stations 66 and 67 by LN are given by the following table

Stn. No.	S_k^2	$\sigma^2_{F,ML}/S_k^2$	K_{S100} MO	ML
66	1.4175	2.73	5.72	8.67
67	2.4943	1.61	7.31	8.58

"Growing" Samples

Hydrologic data series, such as streamflow and precipitation records, are extended with each new year of record. Hydrologists have been concerned with whether predictions based on past records will reasonably agree with future predictions from a larger data base. The problem of outliers has been discussed in the foregoing sections, and suitable methods to deal with samples containing outliers are suggested. In this section, the "constancy" of a "hydrologic population" is tested by comparing σ_F^2 by a shape fitting method to the sample variance of hydrologic data as the size of data increased. For this purpose 23 stations covering a wide range of S_k^2 were picked at random from the 67 stations shown in Table 6.1 and the first 20, 30 and 40 years of data of each of the stations were fit to LN by ML method. The values of $\sigma^2_{F,ML}$ were evaluated for each size sample and the variance ratio σ_F^2/S_k^2 was obtained by dividing $\sigma^2_{F,ML}$ by S_k^2 of the corresponding sample.

Table 6.10 summarizes the results with the first 20, 30 and 40 years of data and the entire available record, which ranged from 55 to 104 years. Table 6.10 shows that, in general, for all sample sizes (20 and above) the variance ratios are either (i) close to unity, or (ii) all larger than 1.0 or all less than 1.0. (The sudden decrease shown in the variance ratios of stations 44, 56, 64 and 65 was due to the presence of higher outliers in the added data). The "Best Fit" criterion applied to these "growing" samples would readily indicate that the "population" is constant for each station regardless of the sample size. Once the "best" PDF for the station is determined the minor changes that occur in variance of fitted PDF as the sample size grew introduce only minor changes in predictions. With regard to the occurrence of higher outliers in future data, since the shape fitting methods have a "correcting" effect upon the fit, predictions based on a larger sample with outliers will be, in general, consistent with those based on a smaller sample without the outlier. However, hydrologists should note that the influence of lower extreme values in multiple number, if any, in the future data is to make $\sigma^2_{F,ML/LS/MCS}/S^2_k$ larger than 1.0 for LN which would result in over predictions by ML/LS/MCS methods when compared to the predictions by MO.

Table 6.10. Frequency Analysis with 'Growing Samples' -

Lognormal Analysis by Maximum Likelihood

	σ_F^2/S_K^2 AT SAMPLE SIZE			
STN NO	20	30	40	A*
11	1.466	1.441	1.368	1.286(77)
12	0.977	1.011	0.953	0.911(73)
13	0.850	0.937	1.128	1.202(75)
19	0.871	0.993	1.013	1.012(77)
23	0.993	0.935	0.881	1.007(71)
24	1.278	1.018	0.937	1.019(57)
30	1.138	1.062	1.042	0.990(59)
31	2.011	1.871	1.231	1.466(80)
33	0.885	0.978	1.076	0.942(82)
40	1.638	2.005	1.720	1.200(76)
41	1.329	1.179	1.037	0.917(81)
42	0.748	0.920	0.882	0.854(72)
44	1.102	1.034	1.195	0.729(72)
48	1.325	1.190	1.219	1.203(69)
49	1.107	1.400	1.401	1.118(104)
55	2.280	2.419	3.104	2.332(64)
56	1.053	1.016	0.753	0.689(56)
58	0.863	0.742	0.720	0.912(68)
60	0.849	0.977	1.051	1.048(62)
64	1.081	0.455	0.469	0.756(62)
65	1.589	1.882	1.895	0.790(56)
66	3.767	3.302	4.199	2.730(64)
67	1.498	1.714	2.353	1.610(55)

A* - AVAILABLE SAMPLE SIZE. THE FIGURES IN THE PARENTHESES INDICATE THE SAMPLE SIZE

CHAPTER VII

CONCLUSIONS AND RECOMMENDATIONS

Summary of Results

The objective of this study was to investigate potential criteria for selecting the probability density function (PDF) of best fit for hydrologic data. The major steps were the examination of various statistical methods for parameter estimation, the determination of differences between PDF's (lognormal (LN) etc. were used as examples), particularly how the characteristic shapes of the PDF's depend on the value of population variance, the examination of the discrepancies between the sample moments and the moments of the fitted PDF, and the identification of strengths and weaknesses of some potential "best fit" criteria. The conclusions from this study are:

1. The three PDF's examined in this investigation (LN, GA and GU) have similar shapes for certain values of the population variance but differ greatly for other values of variance. The comparison was facilitated by basing all analyses on random variables made dimensionless by dividing each by its mean value. Specific conclusions regarding the comparison of the three PDF's are as follows:

(a) Modified GU data (i.e., GU data with negative variates discarded) were found to readily fit a LN PDF up to a variance (σ_k^2) of about 0.2 and a GA PDF in the variance range of 0.2 to 0.6 (at $\sigma_k^2 > 0.6$ the negative tail of GU has more than 5% of the total area and is probably not useful for hydrologic analysis). Thus over this range of variance, use of the GU distribution is redundant.

(b) At a given variance (σ_K^2) the lognormal distribution has a greater skewness (positive) than gamma which makes the two distributions different. For larger return periods (50 years and above) and $\sigma_K^2 < 1.0$, lognormal gives larger predictions than gamma at a given variance.

2. When fitting a sample to a PDF, the method selected for parameter estimation ignores certain characteristics of the sample while it emphasizes others. The method of moments (MO) ignores the overall form of the sample distribution (histogram) while fitting the first two moments of the distribution. On the other hand, the method of maximum likelihood (ML) and the method of least squares (LS), each in their own way, fit only the form or shape of the sample and ignore the sample moments. Specific conclusions relating to the use of these methods and PDF's are as follows:

(a) If a lognormal PDF is fit by a shape fitting method to a data sample having a gamma distribution, the variance of the fitted distribution, on the average, will be larger than the sample variance. This will result in over-prediction by shape fitting methods compared to predictions based on the method of moments.

(b) If a gamma PDF is fit by a shape fitting method to a data sample having a lognormal distribution the variance of the fitted distribution, on the average, will be smaller than the sample variance. This will result in under-prediction by shape-fitting methods compared to predictions based on the method of moments.

3. Noting the results in Conclusion 2, it was concluded that the sample moments will be approximately equal to the moments of the fitted PDF when the correct, or parent, PDF is fitted by a shape fitting method

such as LS or ML. Thus, a criterion for selecting an appropriate probability density function f(x), for fitting a random sample (X_i), of size n is to select that f(x) which makes the statistic

$$\frac{\text{Var } (X_i | f(x;\theta))}{\text{Var } (X_i)}$$

closest to unity, where

$$\text{Var } (X_i) = \sum^{n} (X_i - \bar{X})^2/(n-1)$$

$$\text{Var } (X_i | f(x;\theta)) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x;\theta)dx$$

and the parameters of f(x), θ, are estimated by a shape fitting method such as maximum likelihood or least squares and the mean value, μ, is estimated from

$$\mu = \int_{-\infty}^{\infty} xf(x;\theta)dx$$

and $\bar{X}$ is given by

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

(This criterion is herein referred to as the variance ratio.) Other criteria identified as possible discriminators of PDF's were the statistics of chi-square and Kolomogorov-Smirnov (K-S) goodness-of-fit tests, sum of squared errors of least squares fit, and the range of statistical tolerance limits. Examination of these criteria produced the following conclusions:

(a) The results of numerical experiments indicated that the criterion based on tolerance limits was not useful to identify the parent PDF's of data samples.

(b) The statistics of chi-square and K-S goodness-of-fit tests were found to identify parent PDF's in a majority of cases although some inconsistencies were observed.

(c) While the chi-square statistic was found to be more effective in identifying LN samples compared to GA samples, the K-S statistic was found to be more effective in identifying GA samples than in identifying LN samples.

(d) The above inconsistencies limit the usefulness of these statistics to discriminate PDF's.

(e) A criterion based on the sum of squared errors of data fitted by least squares was found to be generally less effective (compared to chi-square and K-S test statistics) to discriminate PDF's of samples.

(f) The variance ratio was found to be the most satisfactory criterion for use in identifying the parent PDF.

4. The variance ratio test was applied to 67 real hydrologic samples (annual peak flows) and a 'best fit' to either a LN, GA or GU PDF was determined.

(a) The variance ratio test identified LN or GA as the parent or the 'Best' PDF for 44 (66%) samples. (The GU was never found to be unequivocally superior to both GA and LN due to the reasons explained earlier.)

(b) Nineteen samples were judged to contain outliers and neither the LN nor GA distributions provided acceptable fits for four samples.

(c) For samples with outliers the variance of PDF (LN or GA) fitted by a shape fitting method was invariably found to be much less (less than 90%) of the sample variance.

(d) When the variance ratio test was applied to samples with outliers, LN was generally preferred to GA.

(e) Analysis with real samples showed that if LN was fit by a shape fitting method to samples which contained frequent low values, the variance of the fitted PDF would be significantly larger than the sample variance. This, in turn, resulted in large over-predictions by LN.

5. Analysis of the first 20 values of several real data sets was performed and the PDF of best fit (variance ratio) was determined. Then 10 additional years of data was added and the PDF of best fit again was selected. This was repeated until the entire data series was utilized. The results showed that the variance ratio test was consistent in the selected PDF for all sample sizes from 20 to the total of record.

6. If a positively skewed data sample containing outliers (extremely large observations, being far removed from the trend of the other observations) is fit to a positively skewed PDF by a shape-fitting method, the variance of the fitted distribution, in general, will be less than the sample variance. This will result in under-prediction by shape-fitting methods compared to the method of moments.

7. Errors in predictions introduced in frequency analysis by not choosing the 'best' probability density function are larger when computations are made by the maximum likelihood or least squares method than when the computations are made by the method of moments.

Recommendations

Upon the strength of the above conclusions and observations, comparison of sample variance to the variance of a PDF fitted by ML/LS may be the most powerful method for selecting an appropriate PDF to get a given sample. It is seen that, when the estimates of hydrologic events for various frequencies based on a statistically superior method like maximum likelihood are significantly larger compared to the sample observations the statistic of the 'Best Fit' criterion, σ_F^2 / S^2, is significantly larger than unity. When $\sigma_{F\ ML/LS}^2/S^2$ is close to unity, the fit by moments and the fit by maximum likelihood are approximately equal and (the results) appear commensurate with the sample observations. In general, the numerical value of the statistic $\sigma^2_{F,ML/LS}/S^2$, if not equal to unity, will be indicative of several aspects of the sample. In this study, it was seen that definite trends exist in the variation of $\sigma^2_{F,ML/LS}/S^2$ with regard to what would happen if data of a specific PDF are fit to other PDF's and when anomalous observations like outliers are present in data samples. Further studies may be made to establish such trends in the variation of $\sigma^2_{F,ML/LS}/S^2$ when the distribution of sample and the hypothesized distribution differ.

An engineering hydrologist may adopt the following sequence of steps to obtain the best possible estimates of flood peaks for different frequencies from a given record.

1. Assume that LN and GA are two possible best PDF's applicable to flood frequencies.

2. Transform data into dimensionless variables, K_i, by using the equation

$$K_i = Q_i/\bar{Q}$$

in which Q_i is the ith data item and $\bar{Q}$ is the sample mean. Compute the variance of dimensionless data, S_K^2.

3. (i) Compute the variance contributed by the highest data item, Km=1, by Equation 6.1. Determine if Km=1 is an outlier by Figure 6.3.

(ii) If $S_K^2 \geq 1.0$ examine the histogram of data for possible exponential type of distribution. If the sample has many data items close to zero use both GA and LN. Otherwise, GA may not be applicable.

4. Fit the sample to LN and GA by ML (and/or WLS). Compute the ratio σ_F^2/S_K^2 (where σ_F^2 is the variance of the fitted PDF calculated by using the estimated parameters) in each case.

(i) If $\sigma_F^2/S_K^2 \simeq 1.0$ (deviation not more than $\pm$ 0.15) for one of the PDF's that PDF may be regarded as the 'best' PDF applicable to data. The results given by the 'best' PDF thus determined are the best possible estimates of flood peaks for different frequencies (MO and ML/WLS will give approximately equal estimates in this case).

(ii) If $\sigma_F^2/S_K^2 < 1.0$ for both LN and GA the sample contained, possibly, an outlier. This can be verified from the results

of Step 3 (ii). In this case, the PDF for which σ_F^2/S_K^2 has a higher value may be selected as the 'best' PDF. (MO estimates will be higher than ML/WLS estimates in this case. MO estimates may be used where a conservative result is needed).

(iii) If σ_F^2/S_K^2 is greater than 1.0 for LN but less than 1.0 for GA the sample has probably an outlier and many low valued data (close zero) items. GA is the best PDF in this case. MO estimates may be used where a conservative result is needed.

(iv) If σ_F^2/S_K^2 is greater than 1.0 for both LN and GA the data do not fit the selected PDF's. Examine the distribution of the sample (i.e., the sample histogram). For example, the sample probably has a negative skew while LN and GA are positively skewed. Select a PDF, other than LN and GA, which may best fit the sample.

Needed Follow-up Research

The logarithmic probability distributions:

Taking a cue from the lognormal distribution hydrologists have introduced many log probability distributions like log-Pearson type III, log-gamma and the log-Gumbel. As far as the logarithmic probability distributions are concerned, to date, complete knowledge exists only on lognormal PDF in the literature of probability and

statistics. For example, if y, the logarithm of the random variable x, is normally distributed with mean μ_y and variance σ_y^2 the random variable x is distributed as

$$f(x) = \frac{1}{x\sigma_y\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x - \mu_y}{\sigma_y}\right)^2}, \quad x>0$$

$$= 0, \text{ otherwise.}$$

The above distribution is called the lognormal distribution of random variable x. f(x) is always positively skewed. (Figure 3.1 gives the shapes LS PDF for various values of σ_k^2, where $k - x/\bar{x}$). Thus, the form of the distribution of x is known when lnx is normally distributed. Further, the ML estimates $(\hat{\mu}_y, \hat{\sigma}_y)$ of (μ_y, σ_y) of f(x) from a sample of size n are given by

$$\hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n} \ln x_i$$

$$\text{and } \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \ln x_i - (\hat{\mu}y)^2}$$

The above equations also represent the MO estimates of (μ_y, σ_y) for f(y) which is a normal distribution with parameters (μ_y, σ_y) and $y_i - \ln x_i$. To sum up, in case of lognormal distribution (i) If logarithm of a random variable (r.v.) is normally distributed the r.v. has a distribution which is positively skewed, and (ii) If

logarithms of the sample of a r.v. are fitted to a normal distribution by MO, computationally, the sample is fitted to the untransformed distribution, i.e., f(x), by ML method, hence the shape of the sample is fitted.

Similar information does not exist in case of distributions like log-Pearson type III, log-gamma and log-Gumbel. The form of the distribution of the random variable and the statistical moments like μ_x, σ_x^2, γ_{ix}, etc. (i.e., mean, variance, skewness coefficient, etc.) of these distributions are not known. Hence the MO method of estimating parameters of various log-distributions, in general, cannot be known. The present numerical procedures with respect to log-distribution consist of transforming data into logarithms and fitting the transformed data to different PDF's by MO. Since the form of f(x) is not known it cannot be established what these empirical methods really accomplish, i.e., it is not known to what kind of distributions the data are fitted to the untransformed distribution, f(x), of the random variable. The studies made with LN PDF in this work show that to evaluate the results given by a log-PDF the form of f(x) and the statistical method by which data are fit to f(x) should be known. The statistician-hydrologists should make an attempt to derive/establish the form of f(x) and its statistical properties for each of the logarithmic distributions being used.

In case of log-PDF's the 'best-fit' criterion of the present study will be useful to establish

(i) Whether logarithms of hydrologic data can fit distributions

like GA, GU, etc., and

(ii) if logarithms of hydrologic data fit GA, GU, etc., how do the estimates of predictions given by log-PDF's compare with the predictions given by other PDF's for the same data sample.

The Gumbel Distribution

Although the Gumbel distribution has been very popular in certain quarters (Weather Bureau, for example), it appears the negative tail of GU did not draw sufficient attention. In hydrology, only positive data are fitted to GU PDF. An examination of the GU simulated data in this work showed that the statistical properties of the modified GU data (i.e., the samples with the negative data discarded) differed from the population properties when the variance was large. In general, the GU did not fit well the modified GU data at $\sigma_k^2 > 0.6$, hence it is not advisable to fit GU to samples with such large variance. However, procedures are available in statistics to derive truncated distributions by which the positive area of GU may be made equal to 1.0. Figure 3.3 shows that the truncated GU has shapes which are greatly different from LN or GA and they may fit well the samples with many values near zero. Statistician hydrologists should make an attempt to derive the truncated GU distributions.

Some samples may contain data items which are equal to zero (for example, a sample of low flows). For most distributions like LN and GA (at $\sigma_k^2 < 1.0$) both PDF and CDF are zero at k (or x) = 0.0, thus theoretically many distributions are not applicable to samples with

zero values. A search should be made for PDF's which may accommodate zero data items. The truncated GU discussed in the foregoing sub-heading holds a good promise for data with zero values since f(x) can be positive at x = 0 for such a distribution.

APPENDIX A

FREQUENCY ANALYSIS BY THE METHOD OF NON-LINEAR LEAST SQUARES

Section I: Some Theoretical Concepts

If the model given by Equation 2-15a (A.1 below) can not be converted into a form linear in the parameters $\underset{\sim}{\theta}$, the model is said to be intrinsically non-linear (see Chapter II).

$$y_i = f(x_i, \underset{\sim}{\theta}) + e_i \qquad i=1,2,\ldots,n \qquad (A.1)$$

The three PDF's discussed in Chapter III, namely, LN, GA, and GU, are examples of such models. The normal equations for such models (see Equations 2.19) also will be non-linear and obtaining a solution can be extremely difficult. Problems associated with finding solutions to Equations 2.19 of a non-linear system are discussed by Levenberg (1944), Snyder (1962), Hartley (1961), Marquardt (1963), Hartley and Booker (1965), Decoursey and Snyder (1969), Draper and Smith (1966) and Bard (1974), among others.

In most approaches a solution is obtained by linearization of the given model by Taylor series (some associated problems will be dealt with subsequently). The linearized equations are then used recursively until in successive iterations a specified accuracy is reached.

By using Taylor expansion, the linearized normal equations for a nonlinear system may be derived as follows:

The residual e_i in Equation A.1 may be written as

$$e_i(\underset{\sim}{\theta}) = y_i - f(x_i, \underset{\sim}{\theta}) \tag{A.2}$$

For clarity weight w_i (see Chapter II) assigned to e_i^2 will be assumed as unity. (In Section II of this appendix, linearized normal equations are derived retaining w_i for a two parameter nonlinear model.) Choosing as an initial solution $\underset{\sim}{\theta}_o$, some preselected point in $\underset{\sim}{\theta}$-space, at which it is assumed that the sum of squared residuals SSE does not have a stationary value, the first order Taylor expansions of the residuals are taken about $\underset{\sim}{\theta}_o$, giving a set of linear approximations to the residuals,

$$e_i(\underset{\sim}{\theta}) \simeq G_i(\underset{\sim}{\theta}) = e_i(\underset{\sim}{\theta}_o) + \sum_{k=1}^{m} \frac{\partial e_i}{\partial \theta_k} \Delta\theta_k \tag{A.3}$$

where m-number of parameters, and $\Delta\theta_k = \theta_k - \theta_{ko}$, and the partial derivatives are evaluated at $\underset{\sim}{\theta}_o$. Now the standard least squares method consists of minimizing

$$H(\underset{\sim}{\theta}) = \sum_{i-1}^{n} G_i^2 \tag{A.3}$$

by setting the partial derivatives of H with respect to the parameters equal to zero, yielding,

$$\frac{1}{2}\frac{\partial H}{\partial \theta_1} = S_{\theta_1\theta_1}\Delta\theta_1 + S_{\theta_1\theta_2}\Delta\theta_2 + \ldots + S_{\theta_1\theta_m}\Delta\theta_m + R_{\theta_1} = 0$$

$$\frac{1}{2}\frac{\partial H}{\partial \theta_2} = S_{\theta_2\theta_1}\Delta\theta_1 + S_{\theta_2\theta_2}\Delta\theta_2 + \ldots + S_{\theta_2\theta_m}\Delta\theta_m + R_{\theta_2} = 0$$

. . . . . .

$$\frac{1}{2}\frac{\partial H}{\partial \theta_m} = S_{\theta_m \theta_1}\,\Delta\theta_1 + S_{\theta_m \theta_2}\,\Delta\theta_2 + \ldots + S_{\theta_m \theta_m}\,\Delta\theta m + R_{\theta_m} = 0, \qquad (A.4)$$

where,

$$S_{kj} = \sum_{i=1}^{n} \frac{\partial e_i}{\partial \theta_\kappa}\frac{\partial e_i}{\partial \theta_j} \quad \begin{matrix} k=1,2\ldots m \\ j=1,2\ldots m \end{matrix} \quad \text{and} \quad R_j = \sum_{i=1}^{n} \frac{\partial e_i}{\partial \theta_j} e_i \quad j = 1,2\ldots m$$

Equations A.4 are a set of m linear equations in the m unknowns $\Delta\underset{\sim}{\theta}$ and may be solved for $\Delta\underset{\sim}{\theta}$ by standard procedures. Thus, starting from an initial parameter estimate $\underset{\sim}{\theta}_o$, one computes from equations A.4 a set of values of $\Delta\underset{\sim}{\theta}^1$, and sets

$$\underset{\sim}{\theta}_o^1 = \underset{\sim}{\theta}_o + \Delta\underset{\sim}{\theta}^1$$

$\underset{\sim}{\theta}_o^1$ is then used as the new value of $\underset{\sim}{\theta}_o$ in Equations A.3, and the procedure is repeated. After n cycles of the above procedure, one obtains

$$\underset{\sim}{\theta}_o^1 = \underset{\sim}{\theta}_o^{n-1} + \Delta\underset{\sim}{\theta}^n \qquad (A.5)$$

The procedure is terminated when the length of the vector $\Delta\underset{\sim}{\theta}^n$ falls below some preselected value.

The problems associated with linearization procedures are as follows:

1. Convergence may be very slow; that is, very large number of iterations may be required before the solution conver-

ges even though the sum of squares $SSE(\underset{\sim}{\theta}_j)$ (see Equation 2.18) may decrease consistently as j increases.

2. The solutions may oscillate widely, continually reversing direction, and often increasing as well as decreasing the sum of squares. However, the solution may converge eventually.
3. Convergence may not occur at all, even diverge, so that the sum of squares increases iteration after iteration without bound.

One way to combat the foregoing deficiencies is to amend the correction vector $\Delta\underset{\sim}{\theta}^n$ in equation A.5. A program written by Booth and Peterson (1958) amends the correction vector $\Delta\underset{\sim}{\theta}^n$ in A.5 by halving it if

$$SSE(\underset{\sim}{\theta}_o^n) > SSE(\underset{\sim}{\theta}_o^{n-1})$$

or doubling it if

$$SSE(\underset{\sim}{\theta}_o^n) < SSE(\underset{\sim}{\theta}_o^{n-1})$$

This halving and/or doubling process is continued until three points between $\underset{\sim}{\theta}_o^n$ and $\underset{\sim}{\theta}_o^{n-1}$ are found which include between them a local minimum of $SSE(\underset{\sim}{\theta})$. A quadratic interpolation is used to locate the minimum, and the iterative cycle begins again. In theory, this method always converges (Hartley, 1961).

As far back as 1944, Levenberg conceived the idea of "damping" the absolute values of $\Delta\underset{\sim}{\theta}$ as a solution to divergence of $SSE(\underset{\sim}{\theta})$. He replaced Equation A·4 by the related objective function

$$\overline{H}(\underset{\sim}{\theta}) = \xi H(\underset{\sim}{\theta}) + \sum_{k=1}^{m} a_k (\Delta\theta_k)^2 \qquad (A\cdot6)$$

where a_k, k=1,2,...m are a system of positive constants or weighting factors expressing the relative importance of damping the different increments $\Delta\theta_k$, and ξ is a positive quantity expressing the relative importance of the residuals in this minimizing process. While the values of weights ξ and a_i are arbitrary and may be adapted to the requirements of the problem at hand, a particularly effective set of values for a large class of problems has been found to be given by

$$a_k = \sum_{i=1}^{n} \left[\frac{\partial e_i}{\partial \theta_k}\right]^2 \qquad (A\cdot7)$$

and ξ is completely arbitrary and may be varied from iteration cycle to iteration cycle. Minimization of Equation A·6 by substituting for a_k given by Equation A·7, will yield the damped normal equations which are similar to equations given by A·4 except that the principal diagonal elements, $S_{\theta_k\theta_k}\Delta\theta_k$, are multiplied by the factor $(1 + 1/\xi)$. To improve convergence using this device, Grant (1973) set ξ=5.0 initially and doubled it after each iteration. In the work reported herein, the same procedure was used with apparent success.

An alternative to linearization procedure is the method of steepest

descent. This method consists of finding the minimum of a nonlinear function by choosing the values of $\Delta\underset{\sim}{\theta}^n$ in equation A-5 proportional to the negative of the gradient of $SSE(\underset{\sim}{\theta})$ at the point $\underset{\sim}{\theta}_1^{n-1}$. This is a technique of great value in experimental work for finding stationary values of response surfaces. A full description of the method is given in'Design and Analysis of Industrial Experiments' (Davies, 1954).

A method developed by Marquardt (1963) represents a compromise between the linearization method and steepest descent method and appears to combine the best features of both while avoiding their most serious limitations. Essentially, Marquardt's method also consists of determining an appropriate value of ξ (in Levenberg's equations) at each iteration to give rapid convergence of the iterative procedure. His algorithm selects ξ so that the resulting correction vector $\Delta\underset{\sim}{\theta}^n$ is an optimum interpolation between the correction vector obtained by the linearization method and the correction vector obtained by the steepest descent method. Thus, it may be viewed as a refinement of the method of Levenberg. The method almost always converges and does not "slow down" as the steepest descent method often does. However, in the words of Draper and Smith (1966), no method can be called "best" for all nonlinear problems. All methods will work perfectly well on many practical problems which do not violate the limitations of the methods. In general, given a particular method, a problem can usually be constructed to defeat it. Alternatively, given a particular problem and a suggested method, ad hoc modifications can often

provide quicker convergance than an alternative method. Because of the success of relatively simple method of Levenberg in the problems considered in this work, other methods mentioned in the foregoing paragraphs were not examined.

Confidence Regions

A difficulty with the methods of estimation presented thus far is that they do not tell us how close to the parameter $\underset{\sim}{\theta}$ the estimate $\hat{\underset{\sim}{\theta}}$ is likely to be. Since the estimate $\hat{\underset{\sim}{\theta}}$ of the parameter $\underset{\sim}{\theta}$ is based on a random observation $(\underset{\sim}{x}_n, \underset{\sim}{y}_n)$ $\hat{\underset{\sim}{\theta}}$ is a random variable and is not expected to agree exactly with the true, but unknown, value of $\underset{\sim}{\theta}$. One is often interested in estimating the probable region, at some confidence level, within which $\underset{\sim}{\theta}$ lies. More specifically, one seeks a 2m-dimensional region R_λ in $\underset{\sim}{\theta} \times \hat{\underset{\sim}{\theta}}$ space for which the probability of a given point $(\underset{\sim}{\theta}, \hat{\underset{\sim}{\theta}})$ falling within this region is equal to λ (Grant, 1973). It may be recalled that m is the number of components in the parameter vector $\underset{\sim}{\theta}$. The number λ is called the confidence coefficient.

While a general solution leading to a viable computation procedure to construct exact confidence regions is not possible, a computational method, devised by Grant (1973) for a 2-parameter nonlinear distribution based on approaches by Halperin (1962) and Hartley (1964) will be presented here. The solution to this problem commences from the known results from linear least-square theory. If the model given by Equation A 1 is linear it may be written as

$$\underset{\sim}{y} = \underset{\sim\sim}{X}\underset{\sim}{\theta} + \underset{\sim}{e} \tag{A.8}$$

where e is assumed to be a set of N independent errors from $N(0,\sigma^2)$ with σ unknown and the matrix $\underset{\sim}{X}$ is assumed of rank m. The least squares estimate of $\underset{\sim}{\theta}$ is then

$$\underset{\sim}{\hat{\theta}} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\,\underset{\sim}{X}'\underset{\sim}{y} \tag{A.9}$$

which is a best linear unbiased estimate and represents a set of m statistics jointly sufficient for the estimation of $\underset{\sim}{\theta}$. This estimate may be used for the construction of a joint confidence region for some or all of the $\underset{\sim}{\theta}$. The method consists of decomposition of sum of squares errors $\underset{\sim}{e}'\underset{\sim}{e}$ into 'regression' and 'residual' components given by

$$\underset{\sim}{e}'\underset{\sim}{e} = \text{reg}(\underset{\sim}{e}) + \text{res}(\underset{\sim}{e}) \tag{A.10}$$

where the first component

$$\text{reg}(\underset{\sim}{e}) = (\underset{\sim}{X}'\underset{\sim}{e})'\,(\underset{\sim}{X}'\underset{\sim}{X})^{-1}(\underset{\sim}{X}'\underset{\sim}{e}) \tag{A.11}$$

is of rank m and is distributed as $\sigma^2\chi^2$ for m degrees of freedom, while the second component

$$\text{res}(\underset{\sim}{e}) = \underset{\sim}{e}'\underset{\sim}{e} - \text{reg}(\underset{\sim}{e}) \tag{A.12}$$

has rank (N-m) and is independently distributed as $\sigma^2\chi^2$ for (N-m) degrees of freedom. Thus, the ratio of the two components $\text{reg}(\underset{\sim}{e})/\text{res}(\underset{\sim}{e})$ is distributed as Snedecor's "F" statistic with m and N-m degrees of freedom.

The 100λ percent confidence region for $\underset{\sim}{\theta}$ is given by

$$R_\lambda = \left[\underset{\sim}{\theta}: \frac{\text{reg}(\underset{\sim}{e})}{\text{res}(\underset{\sim}{e})} \leq \frac{m}{N-m} F(100\lambda;\ m,\ N-m) \right] \tag{A.13}$$

If the model given by Equation A.1 is nonlinear, equation A.13 may still be used to determine 100λ percent confidence region for $\underset{\sim}{\theta}$. However, in this case the form of the decomposition of $\underset{\sim}{e}'\underset{\sim}{e}$ into the components $\text{reg}(\underset{\sim}{e})$ and $\text{res}(\underset{\sim}{e})$ is no longer obvious. Hartley (1964) has proposed a decomposition, based upon the use of Lagrange's interpolation formulae, to obtain a quasilinearization of the regression function f. Halperin (1962) has proposed a decomposition of the form

$$\text{reg}(\underset{\sim}{e}) = (\underset{\sim}{B}'\underset{\sim}{e})' (\underset{\sim}{B}'\underset{\sim}{B})^{-1}(\underset{\sim}{B}'\underset{\sim}{e}) \tag{A.14a}$$

where

$$\underset{\sim}{B} = \frac{\partial f(x_i;\underset{\sim}{\theta})}{\partial \theta k} = (B_{ij}) \tag{A.14b}$$

These equations are immediately applicable to regression of the system defined by equation 2.15a when $f(x;\ \underset{\sim}{\theta})$ is nonlinear.

For a two-parameter nonlinear system, i.e., when m=2, using Equations A.14, Grant (1973) computes the values of the statistic F by

$$F = \frac{N-m}{m} \frac{\text{reg}(\underset{\sim}{e})}{\text{res}(\underset{\sim}{e})} \tag{A.15}$$

at fixed points of a grid surrounding the least squares estimates of θ_1, θ_2. This makes it possible, at each point of the grid, to compute the cumulative probability at the corresponding value of F; that is,

$$\zeta = \int_0^F dF$$

The values of ζ are then plotted in the θ_1-θ_2 plane at locations corresponding to the values of (θ_1,θ_2) for which each ζ was derived. The various confidence regions are then constructed by sketching a closed curve (using interpolation where necessary) passing through the value ζ for which the confidence region was desired. Confidence regions constructed using Grant's method are shown in Chapter V.

Tolerance Limits

In the foregoing section it has been found that confidence regions may be determined so that the limits of the region will cover a population parameter $\underset{\sim}{\theta}$ with certain confidence, i.e., a certain proportion of the time. Sometimes it is desirable to obtain limits which will cover a fixed-proportion of the population distribution with a specified confidence. In other words, if some "best" estimate $\hat{\underset{\sim}{\theta}}$ of $\underset{\sim}{\theta}$ is deduced from an observation $(\underset{\sim}{x}_n,\underset{\sim}{y}_n)$, it is desired to determine a value $\upsilon_{\gamma u,\lambda}$ which is a function of $(\underset{\sim}{x}_n,\underset{\sim}{y}_n)$ such that γ proportion of the population will be below $\upsilon_{\gamma u,\lambda}$ with 100λ percent certainty. Here λ has the same meaning as defined earlier and the value $\upsilon_{\gamma u}$ is such that

$$\int_{-\infty}^{\upsilon_{\gamma u,\lambda}} p(x;\hat{\underset{\sim}{\theta}}_{u\lambda})\, dx = \gamma \qquad (A.16)$$

where p is a probability frequency function of the random variable X and $\hat{\underset{\sim}{\theta}}_{u,\lambda}$ is the estimated value of parameter at 100λ% upper confidence limit. Also, $\upsilon_{\gamma u,\lambda}$ computed from equation A.16 may be expected to vary from observation to observation, hence a random variable. However, from the definition of confidence limits of $\underset{\sim}{\theta}$ the random variable $\upsilon_{\gamma u,\lambda}$ has the property that for large number of observations the inequality

$$(\upsilon_{\gamma u,\lambda})_{true} \leq \upsilon_{\gamma u,\lambda}$$

where $(\upsilon_{\gamma u,\lambda})_{true}$ is the actual but unknown 100γ percentile point of the population, may be expected to be true at least for 100λ% of observations. In this case $\upsilon_{\gamma u,\lambda}$ is called an upper tolerance limit. Similarly a lower tolerance limit $\upsilon_{\gamma L,\lambda}$ may be defined as a random variable above which γ proportion of population will be, with 100λ percent certainty. Methods are available to compute the tolerance limits if the observation $(\underset{\sim}{x}_n, \underset{\sim}{y}_n)$ belongs to a normal population Natrella (1963), Bowker and Lieberman (1972). To get an upper tolerance limit in a general case a confidence region R_λ for the population parameters $\underset{\sim}{\theta}$ is first constructed. Then search is made in R_λ for a parameter value $\hat{\underset{\sim}{\theta}}_\lambda$ which will maximize $\upsilon_{\gamma u,\lambda}$ (see Equation A.16). Thus, one may immediately define an upper one sided tolerance limit $\upsilon_{\gamma u,\lambda}$ to be the random variable

$$\upsilon_{\gamma u,\lambda} = \max_{\hat{\underset{\sim}{\theta}} \in R_\lambda} \{\upsilon;\ \int_{-\infty}^{\upsilon} p(x,\hat{\underset{\sim}{\theta}})dx=\gamma\} \qquad (A.17)$$

where $\hat{\underset{\sim}{\theta}}$ is any point in R_λ.

Since by definition of R_λ, $\underset{\sim}{\theta}$ is included in 100λ% of the R_λ computed

from a large number of samples, for at least 100λ% of observations the inequality

$$(\upsilon_{\gamma u,\lambda})_{true} \leq \upsilon_{\gamma u,\lambda}$$

holds. The above inequality is the definition of an upper one-sided tolerance limit.

Section II. Fitting Probability Density Functions to Data Histograms by Least Squares

One of the methods of verifying a proposed probabilistic model is to arrange the observed data in the form of a histogram, then superimpose the proposed PDF on the data histogram and compare the two shapes (Benjamin and Cornell, 1970). Such visual comparison permits an immediate assessment of the proximity of observed and predicted results. In fact, the well known χ^2-goodness of fit test is based on the sum of squared errors (modified) between frequency of each class of data histogram and the average of the proposed distribution for that class. Conversely, one may intuitively feel that one efficient way of estimating parameters of the proposed probabilistic model would be to organize the given set of data into a frequency histogram and optimize the parameters of the distribution to achieve a minimum value of the sum of squared differences between the frequency of each class and the average of the distribution for that class. Initial work by Snyder (1972) and Snyder and Wallace (1974) has shown that such a procedure is mathematically tractable. The procedure commences with

organizing the data sample into a frequency histogram of an appropriate number of class intervals. The distribution function is also expressed in the frequency form for the same grouping (see Section III, Figure A.1). The difference between the observed frequency of a given class and the frequency indicated by the distribution function for the same class is defined as the class error (see Section III, Figure A.2). The least squares procedures described in Section I can now be applied to estimate the parameters (α,β) of the distribution function so that the sum of the squares of the class errors is a minimum.

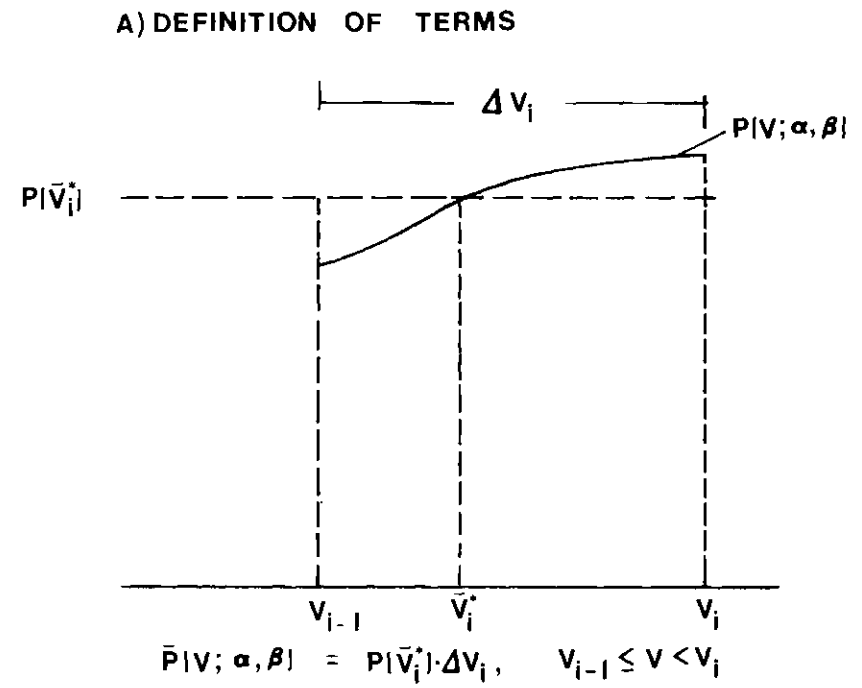
For the ith class interval of the histogram, let h_i be the observed frequency, $\bar{p}(v_i;\underset{\sim}{\theta})$ be the frequency indicated by the distribution function and e_i be the class error. By analogy with Equation A.1, one writes

$$h_i = \bar{p}(v_i;\underset{\sim}{\theta}) + e_i \tag{A.18}$$

The weighted sum of squared errors (see Equation 2.18) for the frequency histogram of N class intervals is given by

$$\text{SSE } (\underset{\sim}{\theta}) = \sum_{i=1}^{N} w_i \left[h_i - \bar{p}(v_i;\underset{\sim}{\theta}) \right]^2 \tag{A.19}$$

While many weighting functions may be chosen for w_i, one natural choice for w_i, motivated by the method of minimum chi-square, appears to be the reciprocal of the density function itself raised to some power (Grant, 1973). That is,



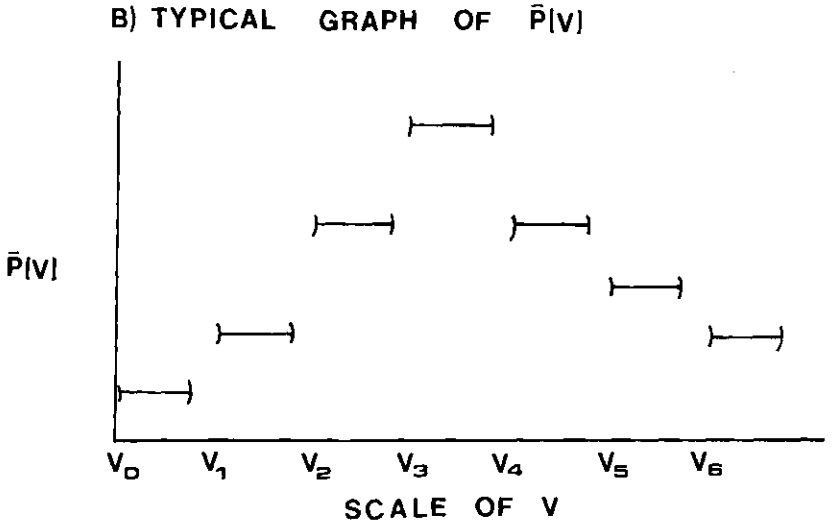


Figure A.1 The Finite Form of the Probability Density Function

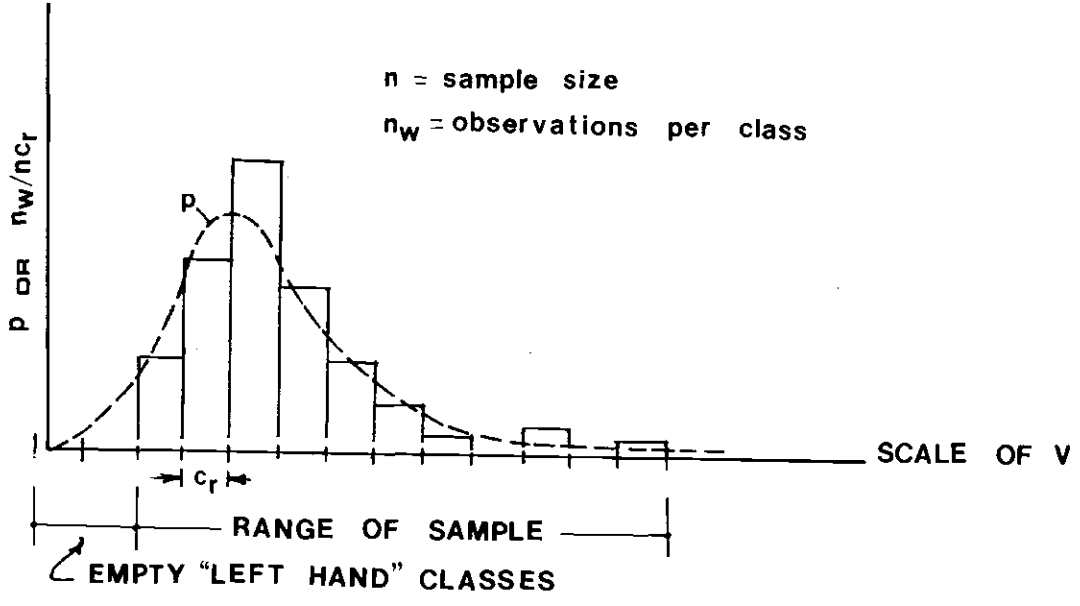


Figure A.2 Elements of a Histogram

$$w_i = \bar{p}(\upsilon_i;\theta)^{-\phi} \tag{A.20}$$

To develop normal equations for use in the method of least squares (weighted) one writes

$$\frac{\partial(SSE)}{\partial\theta_j} = -2 \sum_{i=1}^{N} w_i \left(h_i - \bar{p}(\upsilon_i;\theta)\right) \frac{\partial \bar{p}(\upsilon_i;\theta)}{\partial\theta_j} = 0 \qquad j=1,2,\ldots,m \tag{A.21}$$

where the terms involving $\frac{\partial w_i}{\partial\theta_j}$ have been ignored (see for example, chi-square minimum method, Cramer (1946), or Kendall and Stuart (1973) for justification). The linearization technique described in Section I of this appendix may be applied to Equations A.21 and normal equations in a solvable form may be obtained. The final form of linearized normal equations for a two parameter distribution are shown in Section III of this Appendix. For a case of unitary weights, (i.e., when $w_i=1$ in Equation A.19) the value of weight exponent ϕ, in Equation A.20 may be set equal to zero. Although the above method is apt to be numerically laborious, the nature of the computations is quite routine and easily programmable on a digital computer. Moreover, the basic procedure would remain the same for any type of distribution function, and so once a computer program has been developed, it would, with only minor modifications be able to fit any function desired. Also, the method of constructing confidence regions outlined in Section I allows the (relatively) easy construction of tolerance limits for any distribution function based upon any sample. The foregoing features, thus may have a particular appeal for estimating statistical parameters by the method of least squares.

The Choice of a Class Interval

The choice of the number of class intervals within the range of the data sample may alter the shape of the data histogram. This leads to the suspicion that the manner of grouping the data into histograms may affect the efficiency of the fitting process by least squares. Sturges (1926) suggests that a reasonable choice of a class interval to use in constructing a histogram from a set of n data points is given by the formula

$$C = \frac{R}{1 + 3.322 \log n} \qquad (A.22)$$

where R is the range of the sample and log means the logarithm to the base 10. To study the influence of the class interval on least squares estimates of population parameters, Grant (1973) modifies the Equation A.22 as

$$Cr = \frac{rR}{1+3.322 \log n} \qquad (A.23)$$

where r is a scaling factor on the class interval suggested by Sturges. Using samples (n=50) consisting of gamma synthetic variates of known population parameters, Grant computes the class intervals of data histograms for r= 0.5, 0.75 and 1.0 by Equation A.23. For a given value of Cr, the histogram was constructed by selecting the centerpoint of the first group by the relation

$$x_1 = Cr\left[\frac{x_{min} + 1/2\ Cr}{Cr}\right] \qquad (A.24)$$

where $[aa]$ indicates the integral portion of aa. The centerpoints of

subsequent groups were calculated from the relation

$$x_i = x_{i-1} + Cr \tag{A.25}$$

The frequency histogram was constructed by determining the number of observations falling into each interval (x_i-0.5 Cr, x_i+0.5 Cr), and then by adding to the lower portion of the histogram the smallest number of empty classes required to contain the point x=0 (see Figure A.2). For 100 (synthetic) samples from a given population, frequency histograms were constructed by the procedure described above and least squares estimates of the population parameters were obtained for the three cases, r=0.5, 0.75, and 1.0. Based on results of data sets from 10 different gamma populations, Grant concluded that the choice of the class width had no appreciable effect upon the results of the fitting. Based on Grant's conclusion the value of r was arbitrarily chosen as 0.5 in this study and historgrams were constructed by Grant's procedure described above.

One might intuitively believe that a good grouping should be the one for which the resulting histogram assumed a relatively smooth or recognizable shape. Nevertheless, the probability law of a histogram (Benjamin and Cornell, 1970) indicates that the "perfect fit" is not a highly likely outcome. The probability of not finding a perfect fit, in fact, tends to unity as the sample size increases. Therefore, the engineer should not be surprised if the data do not compare perfectly with his model.

Section III: Development of Normal Equations in Nonlinear Regression to Estimate Parameters of a Two Parameter Probabilistic Model

The normal equations for a two-parameter probabilistic model fit to a sample histogram by nonlinear least squares are described below.

1. Let $p(x;\alpha,\beta)$ denote the probability density function (PDF) selected. α,β are the two parameters of the distribution. To express the PDF in finite form, one writes

$$\bar{p}(\upsilon_i;\alpha,\beta) = \int_{\upsilon_{i-1}}^{\upsilon_i} p(x;\alpha,\beta)dx, \quad \upsilon_{i-1} \leq x \leq \upsilon_i \tag{A.26}$$

By the mean value theorem for integrals (Taylor, 1955), there exists a number υ^* in $\upsilon_{i-1},\upsilon_i$ such that

$$\bar{p}(\upsilon_i;\alpha,\beta) = p(\upsilon^*;\alpha,\beta)\ \Delta\upsilon_i \tag{A.27}$$

where

$$\Delta\upsilon_i = \upsilon_i - \upsilon_{i-1}$$

Equation A.27 is often written in the more informal form

$$\bar{p}(\upsilon_i;\alpha,\beta) = p(\upsilon;\alpha,\beta)\Delta\upsilon_i \tag{A.28}$$

where Equation A.28 should be understood in the sense of Equation A.27 (see also Figure A.1).

2. Let the sample histogram (Figure A.2) be denoted by the sequence $\left[h_i \right]_{i=1}^{N}$ and let it be assumed that the histogram has been normalized so that

$$\sum_{i=1}^{N} h_i = 1 \tag{A.29}$$

3. By analogy with Equation 2.18, the weighted sum of the residual squares SSE which it is desired to minimize becomes

$$SSE = \sum_{i=1}^{N} (h_i - \bar{p}(\upsilon_i;\alpha,\beta))^2 \, w_i \tag{A.30}$$

and so,

$$\frac{\partial(SSE)}{\partial\alpha} = \sum_{i=1}^{N} \left[\frac{\partial w_i}{\partial\alpha} (h_i - \bar{p}(\upsilon_i;\alpha,\beta))^2 - 2w_i (h_i - \bar{p}(\upsilon_i;\alpha,\beta)) \right.$$

$$\left. X \frac{\partial \bar{p}}{\partial\alpha} (\upsilon_i;\alpha,\beta) \right] = 0 \tag{A.31}$$

and

$$\frac{\partial SSE}{\partial\beta} = \sum_{i=1}^{N} \left[\frac{\partial w_i}{\partial\beta} (h_i - \bar{p}(\upsilon_i;\alpha,\beta))^2 \right.$$

$$\left. -2w_i (h_i - \bar{p}(\upsilon_i;\alpha,\beta)) \frac{\partial \bar{p}}{\partial\beta} (\upsilon_i;\alpha,\beta) \right] = 0 \tag{A.32}$$

Assuming that the terms involving $\frac{\partial w_i}{\partial\alpha}$ and $\frac{\partial w_i}{\partial\beta}$ are negligible in comparison to other terms in Equations A.31 and A.32, and for the particular choice of weights

$$w_i = \bar{p}(v_i;\alpha,\beta)^{-\phi} \tag{A.33}$$

Equations A.31 and A.32 may be written in the form

$$\frac{\partial SSE}{\partial \alpha} = -2 \sum_{i-1}^{N} \bar{p}(v_i;\alpha,\beta)^{-\phi} (h_i - \bar{p}(v_i;\alpha,\beta) \frac{\partial \bar{p}}{\partial \alpha} (v_i;\alpha,\beta)=0 \tag{A.34}$$

and

$$\frac{\partial SSE}{\partial \alpha} = -2 \sum_{i-1}^{N} \bar{p}(v_i;\alpha,\beta)^{-\phi} (h_i - \bar{p}(v_i;\alpha,\beta) \frac{\partial \bar{p}}{\partial \beta} (v_i;\alpha,\beta)=0 \tag{A.35}$$

With the usual Taylor approximation of $\bar{p}$ about the point (α_o,β_o), one obtains the normal equations (see Section I),

$$S_{\alpha\alpha}\Delta\alpha + S_{\alpha\beta}\Delta\beta = R_\alpha \tag{A.36}$$

$$S_{\alpha\beta}\Delta\alpha + S_{\beta\beta}\Delta\beta = R_\beta \tag{A.37}$$

where

$$S_{\alpha\alpha} = \sum_{i=1}^{N} \frac{\partial \bar{p}}{\partial \alpha} (v_i;\alpha_o,\beta_o)^2 (\bar{p}(v_i;\alpha_o,\beta_o))^{-\phi} \tag{A.38}$$

$$S_{\beta\beta} = \sum_{i=1}^{N} \frac{\partial \bar{p}}{\partial \beta} (v_i;\alpha_o,\beta_o)^2 (\bar{p}(v_i;\alpha_o,\beta_o))^{-\phi} \tag{A.39}$$

$$S_{\alpha\beta} = \sum_{i=1}^{N} \frac{\partial \bar{p}}{\partial \alpha} (v_i;\alpha_o,\beta_o) \frac{\partial \bar{p}}{\partial \beta} (v_i;\alpha_o,\beta)(\bar{p}(v_i;\alpha_o,\beta_o))^{-\phi} \tag{A.40}$$

$$R_\alpha = \sum_{i=1}^{N} (h_i - \bar{p}(\upsilon_i;\alpha_o,\beta_o)) \frac{\partial \bar{p}}{\partial \alpha}(\upsilon_i;\alpha_o,\beta_o)(\bar{p}(\upsilon_i;\alpha_o,\beta_o)^{-\phi} \quad (A.41)$$

By applying Levenberg's method Equations A.36 and A.37 may be written as

$$(1 + \frac{1}{\xi}) S_{\alpha\alpha}\Delta\alpha + S_{\alpha\beta}\Delta\beta = R_\alpha \quad (A.43)$$

$$S_{\alpha\beta}\Delta\alpha + (1 + \frac{1}{\xi}) S_{\beta\beta}\Delta\beta = R_\beta \quad (A.44)$$

Appropriate values of ξ were found to be, by trial (Grant,1973)

$$\xi_o = 5. \quad (A.45)$$

$$\xi_n = 2\xi_{n-1} \quad (A.46)$$

Using the above relations for ξ, one seeks the least squares estimates a and b of α and β by repeated solution of equations A.43 and A.44. This iteration is continued until $(\Delta\alpha^2+\Delta\beta^2)^{\frac{1}{2}}$ is less than some preassigned quantity (the value used in this study was 10^{-4}).

The initial estimates a_o and b_o necessary to begin the least squares solution may be obtained by any other estimating methods (see Chapter II). Inthis study, moment estimates were used as initial estimates for gamma and Gumbel and maximum likelihood (ML) estimates were used in case of lognormal since computation of ML estimates was always possible for LN distribution.

The derivatives of LN, GA, and GU distributions with respect to the two parameters involved are shown in Appendices B, C, and D, respectively.

Section IV: Study of Nonlinear LS Method by Numerical Experiments

The statistical properties of nonlinear least squares (denoted as LS henceforth) were studied in detail by simulation experiments. LN, GA, and GU PDF's were examined (see Chapter III for a description of the PDF's). The objectives of the numerical simulation experiments were to answer the following questions:

1. Does the LS technique produce unbiased parameter estimates? How is the bias, if it exists, reflected in the LS estimates of PDF percentiles?
2. What is the efficiency of LS fitting when compared to fitting by ML?
3. Is the inclusion of the weighting factors introduced in Chapter II and the particular choice of weighting function given by Equation A.20 necessary for a satisfactory fit by LS? In general, what is the effect of weight, as proposed in this study, on the PDF percentiles?
5. Is the method of LS more or less stable, i.e., the LS estimates of parameters are approximately equal when the method is applied to samples of different sizes from the same population?

6. Are the class errors defined by Equation A.18 (or more generally by Equation 2.17), in general, normally distributed with mean zero and variance σ^2?

In an attempt to obtain answers to above questions 166 computer simulation runs were made. Table 5.1 presents the values of population parameters of LN, GA and GU PDF's used in simulation runs. Chapter V gives a detailed description of salient features of simulation runs. Except in runs made to answer Question No. 5 above, the sample size was 100 for all samples. Appendix G gives a summary description of each of the simulation runs.

Section IV: Description, Results and Analysis of Simulation Experiments

Study 1: Bias in LS estimates.

To examine whether the LS technique (unweighted, ϕ=0) produces biased estimates of parameters (see Chapter II for a definition of unbiasedness) and PDF percentiles several runs with identical population parameters but with different series of random numbers were studied. (Starting from a given initial (random) number, the computer generates a series of pseudo-random numbers. To eliminate bias, if any, due to the initial number supplied different series of random number initiated by different numbers were used to obtain synthetic variates of PDF's). For each series of runs Table A.1 summarizes the mean parameter estimates ($\bar{A}$ and $\bar{B}$), the ratios

Table A.1. Bias in Least Squares Estimates

(n = 100, ϕ = 0.00)

PDF	RUN SERIES	NO OF SAMPLES	σ_K^2	α	β	FIT	$\bar{A}$	$\bar{A}/\alpha$	$\bar{B}$	$\bar{B}/\beta$	K_{S100}/K_{100}
LN	1LN	125	0.0942	-0.0450	0.3000	LS	-0.044	0.97	0.300	1.00	1.008
						ML	-0.047	1.04	0.301	1.00	1.001
	2LN	125	0.1735	-0.0800	0.4000	LS	-0.083	1.04	0.399	1.00	1.002
						ML	-0.082	1.03	0.399	1.00	0.993
	3LN	100	0.2840	-0.1250	0.5000	LS	-0.125	1.00	0.492	0.98	0.983
						ML	-0.128	1.03	0.499	1.00	0.997
	4LN	100	0.6323	-0.2450	0.7000	LS	-0.241	0.98	0.688	0.98	0.995
						ML	-0.240	0.98	0.694	0.99	1.001
GA	1GA	125	0.3333	3.0000	3.0000	LS	3.242	1.08	3.184	1.06	0.980
						ML	3.106	1.04	3.080	1.03	0.991
	2GA	125	0.2000	5.0000	5.0000	LS	5.340	1.07	5.266	1.05	0.981
						ML	5.146	1.03	5.095	1.02	1.013
	3GA	150	0.1429	7.0000	7.0000	LS	7.273	1.04	7.247	1.04	0.996
						ML	7.077	1.01	7.056	1.01	1.000
	4GA	122	0.1000	10.0000	10.0000	LS	10.496	1.05	10.466	1.05	0.994
						ML	10.219	1.02	10.154	1.02	0.994
GU	1GU	150	0.7311	1.5000	0.6152	LS	1.671	1.11	0.640	1.04	0.930
						MO	1.608	1.07	0.747	1.22	0.989
	2GU	150	0.4112	2.0000	0.7114	LS	2.063	1.03	0.709	1.00	0.985
						MO	2.090	1.05	0.739	1.04	0.987
	3GU	150	0.1828	3.0000	0.8076	LS	3.095	1.03	0.802	0.99	0.984
						MO	3.055	1.02	0.808	1.00	0.992
	4GU	125	0.1028	4.0000	0.8569	LS	4.072	1.02	0.860	1.00	0.988
						MO	4.040	1.01	0.859	1.00	0.995

$\bar{A}/\alpha$ and $\bar{B}/\beta$ (where α and β are the population parameters), the ratios $\bar{K}_{S100}/K_{100}$ (where $\bar{K}_{S100}$, K_{100} are the mean sample prediction and population prediction of 100-year event, respectively).

Lognormal Distribution Table A.1 shows that while the LS estimates of β(i.e., σ_y of the PDF, see Table 3.1) are practically unbiased the LS estimates of α(i.e., μ_y of the PDF, see Table 3.1) are slightly (~3% to 8%) biased. The bias (Table A.1 shows the values of $|\bar{A}/\alpha|$) in the estimates of μ_y does not appear to affect the predictions largely. For the run series 3LN which has the maximum bias (+8%) in its LS estimates of μ_y the estimates of 100-year predictions, $\bar{K}_{S100}$, showed a bias of only 1.7%. Compared to LS estimates, the ML estimates are less biased. Both LS and ML estimates of $\bar{K}_{S100}$ are practically unbiased.

Based on the results of Table A.1, it may be stated that LS method gives practically unbiased estimates of parameters and percentiles (i.e., predictions) for LN.

Gamma Distribution Table A.1 shows that, in general, both ML and LS estimates of GA parameters are positively biased, the bias being in the range of 1% to 4% and 4% to 8% in ML and LS estimates, respectively. Whether the GA variates generated by the particular technique used (see Appendix E) would always lead to such positively biased estimates of parameters was not tested in this study. However, the estimates of

100-year predictions, $\bar{K}_{S100}$, are found to be practically unbiased. A reason for such an occurrence may be that the bias in parameter estimates is not significant enough to cause a bias in percentiles.

Based on the results of this study, it may be stated that LS method gives positively biased estimates of GA parameters, but practically unbiased estimates of GA percentiles.

Gumbel Distribution It may be recalled that the limits of the random variable of Gumbel distribution are $\mp \infty$ and the CDF representing the negative values of the random variable increases with the increase of σ_k^2 (see Table 3.3). This characteristic of Gumbel distribution leads to the generation of negative pseudorandom numbers. To conform with the real world hydrologic data the negative pseudorandom numbers generated were always discarded in this study. The expected incidence of negative variates is about 8% and 3% for the two series of runs represented by 1GU and 2GU, respectively and for the other two series of runs (i.e., 3GU and 4GU) it is negligible. The actual incidence of negative random number generated with simulated samples is summarized in Table A.2. The samples were generated such that they would contain the required number of positive variates.

Table A.2. Gumbel Distribution - Number of Negative Variates Generated with any Sample

σ_K^2	Sample Size (Number of Positive Variates Generated) 25	50	75	100	Run Series
0.7311	0 to 6	2 to 10	3 to 14	1 to 19	1GU
0.4112	0 to 1	0 to 4	0 to 5	0 to 6	2GU
0.1828	0	0	0	0 to 1*	3GU
0.1208	0	0	0	0	4GU

* Extremely Rare (only one case occurred in 125 cases)

Table A·1 shows that the least squares estimates of parameters and predictions are more biased for the 1GU series of runs ($\sigma_k^2 = 0.7311$ and the expected incidence of negative variates is about 8%) than for the other series of runs. One obvious reason for this occurrence appears to be the discarding negative variates. To examine the effect of discarding the negative variates on the statistical characteristics of the samples the mean and variance of sample means ($\bar{M}_S$, $S^2(M_S)$) and the mean and variance of sample variances ($\bar{V}_S$, $S^2(V_S)$) were computed for the three series of runs represented by 1GU, 2GU and 3GU. (see Table A·3). While the mean values of sample means and sample variances are found close to the population values for the series of runs represented by 3GU, the same values of the series of runs represented by 1GU are considerably affected by the removal of negative variates. In case of 1GU runs, the sample means were increased by 9 to 15% and the sample variances were decreased by 7 to 11% by discarding the negative variates. One may argue that the two happenings, namely, the increase in mean and the decrease in variance, may have a compensatory effect on the resulting fit. Indeed this argument has been found to be true in case of moments fit (see the value of $\bar{K}_{S100}/K_{100}$ given by the moments method for 1GU, Table A·1). The increased sample means due to the discard of negative variates resulted in a large positive bias

Table A.3. Statistical Characteristics of Gumbel Samples

($n = 100$, $\mu_K = 1.00$)

σ_K^2	RUN NO	M_S	$S^2(M_S)$	V_S	$S^2(V_S)$
.7311	1GU1	1.0961	0.0053	0.6554	0.0259
	1GU2	1.0993	0.0097	0.5696	0.0245
	1GU3	1.0905	0.0056	0.6484	0.0124
	1GU4	1.1039	0.0054	0.6536	0.0317
	1GU5	1.1457	0.0058	0.6823	0.0221
	1GU6	1.1253	0.0077	0.6776	0.0201
.4112	2GU2	1.0150	0.0051	0.4009	0.0060
	2GU3	1.0093	0.0037	0.3890	0.0052
	2GU4	1.0313	0.0055	0.3979	0.0074
	2GU5	1.0442	0.0032	0.4090	0.0076
	2GU6	1.0026	0.0051	0.3820	0.0097
.1823	3GU1	0.9915	0.0023	0.1762	0.0019
	3GU2	0.9990	0.0022	0.1795	0.0006
	3GU3	0.9916	0.0016	0.1781	0.0009
	3GU4	1.0001	0.0023	0.1844	0.0021
	3GU5	1.0139	0.0015	0.1906	0.0015
	3GU6	0.9952	0.0017	0.1814	0.0010

M_S = Sample Mean V_S = Sample Variance

Number of Samples = 25

(22%) in the moment estimates of parameter β (see Equation D.8 Appendix D) which in turn resulted in a fit whose $\bar{K}_{S100}$ was close to K_{100} (i.e., population value). On the other hand, since LS method fits the proposed PDF to the shap of a sample, fitting a GU PDF to a GU sample with negative variates discarded is equivalent to fitting a PDF to a dismembered population! Consequently, though both LS and MO estimates of parameters were biased for 1GU run series, the effect of the bias was not same on percentiles in both the methods.

The LS estimates of parameters and $\bar{K}_{S100}$ are practically unbiased for run series 2GU through 4GU (σ_k^2 = 0.10 to 0.41).

Summary It is found that the LS method, in general, produces unbiased estimates of percentiles for LN, GA and samples from low variance populations ($\sigma_k^2 \leq 0.4$) of GU. For high variance populations of GU, LS method produces negatively biased percentiles; discarding of negative GU variates generated at high variances causes such a negative bias. Convergence in LS computations was quite rapid and was 100% for lognormal and Gumbel distributions. For most samples the number of iterations required for convergence was found to be 2 to 5 and 3 to 7 for lognormal and Gumbel distributions respectively. The gamma distributions required about 9 to 10 iterations for convergence, on an average for parameter values 7 and below. When α,β = 10, i.e., for 4GA runs, the number of iterations required for convergence was larger than 10 for most samples and some samples did not converge after 50 iterations. At higher values

of gamma parameter, i.e., at low σ_k^2 , the tails of the distribution, particularly the right hand tail, becomes very thin. The derivatives of the probability function are very small for larger values of random variable, thus the resolving power of the least squares method is reduced if the sample contains some observations in the tail. This may be a cause for slow convergence at higher parameter values of gamma distribution. When weights were assigned to tail errors, convergence was found to be rapid (see Study No. 3).

The choice of maximum likelihood estiamtes as the inital parameter guesses may be the main reason for excellent convergence of lognormal distribution.

Study 2: Efficiency of the Least Squares Estimates

If $\hat{\theta}$ and θ^* are (unbiased) estimators of θ, then $\hat{\theta}$ is said to be relatively more efficient than θ^* if the variance of $\hat{\theta}$ is less than the variance of θ^* (Hines and Montgomery, 1972). Based on the above concept of efficiency, the property of efficiency of least squares estimates has been empirically studied as a part of this work. Twenty five samples belonging to the same population were generated for each simulation run. The sample variance of estimates for each of the simulation runs represents a measure of efficiency of the corresponding estimates.

In Table A.4 are presented the ratios of LS to ML variances of parameter and K_{S100} (100-year sample predictions) estimates, S_A^2(LS/ML), S_B^2(LS/ML), and $S^2_{K_{S100}}$ (LS/ML), respectively, for each run of LN and GA run series. For GU runs, the variance of MO and LS estiamtes are compared

Table A.4. Efficiency of Least Squares Estimators - LN, GA PDF's

(n = 100, ϕ = 0.00)

RUN SERIES		RUN NO	S_A^2(LS/ML)	S_B^2(LS/ML)	S_{KS100}^2(LS/ML)
	1LN	1LN1	1.19	3.78	3.77
		1LN2	1.52	2.89	3.07
RU		1LN3	1.04	2.78	2.42
		1LN4	1.24	1.88	1.84
		1LN5	1.25	1.55	1.44
	2LN	2LN1	1.74	1.67	2.30
		2LN2	1.25	2.27	1.90
		2LN3	1.55	2.83	3.47
		2LN4	1.40	2.37	2.35
		2LN5	1.48	2.05	1.75
	3LN	3LN1	2.42	2.48	3.02
		3LN2	2.61	2.77	2.52
		3LN3	1.42	1.36	1.01
		3LN4	1.49	2.06	1.74
	4LN	4LN1	1.36	1.51	1.37
		4LN2	1.66	2.80	4.41
		4LN3	1.45	2.39	2.33
		4LN4	1.73	2.05	2.52
	1GA	1GA1	2.29	1.78	2.29
		1GA2	1.14	1.84	1.60
		1GA3	4.00	4.48	2.19
		1GA4	1.96	1.88	1.84
		1GA5	1.52	1.35	2.17
	2GA	2GA1	1.67	1.73	1.83
		2GA2	1.26	1.43	1.14
		2GA3	1.91	2.32	1.21
		2GA4	1.41	1.51	1.47
		2GA5	2.45	1.86	3.35
	3GA	3GA1	1.12	1.25	1.03
		3GA2	2.04	2.17	1.73
		3GA3	1.98	2.40	1.77
		3GA4	2.46	2.20	2.29
		3GA5	2.05	2.15	1.31
	4GA	4GA1	2.07	2.21	2.37
		4GA2	1.60	1.84	1.54
		4GA3	1.87	1.76	2.03
		4GA4	1.65	1.66	2.39
		4GA5	2.80	2.70	1.31

(Table A.5) since ML results are not available.

Table A.4 shows that the variance of LS (ϕ=0.00) estimates is invariably larger than the variance of ML estimates for LN and GA PDF's. The mean values of LS to ML variance ratios are in the order of 1.55, 2.27 and 2.40 for the estimates of α,β and K_{S100}, respectively for LN. For GA the mean of the variance ratios was about 2.0 for the three estimates.

For GU, Table A.5 shows that the sample variances of parameter estimates by MO are, in general, less than the sample variances of LS estimates. Nevertheless, values of $S^2_{K_{S100}}$ (MO/LS) show that the LS and MO methods are about equal from the efficiency point of view. With the use of weight (see Study No. 3) the LS estimates invariably become more efficient than the MO estimates.

Table A.6 shows the range of magnitudes of variances S_A^2, S_B^2 and $S^2_{K_{S100}}$ for various run series presented in Tables A.4 and A.5.

Summary: ML estimates of parameters and percentiles based on these parameters are, in general, found to be more efficient than the LS estimates for LN and GA. For GU the unweighted LS method did not show any particular superiority over the method of moments as far as efficiency is concerned.

Study 3: Properties of the Weighted Least Squares Method.

The weighted least squares method, as examined in this study may be regarded as a more general case of the Minimum Chi-Square method. With

Table A.5. Efficiency of Least Squares Estimators - GU PDF

(n = 100, ϕ = 0.00)

RUN SERIES	RUN NO	S_A^2 (MO/LS)	S_B^2 (MO/LS)	S_{KS100}^2 (MO/LS)
1GU	1GU1	2.34	0.21	1.95
	1GU2	0.77	0.76	1.07
	1GU3	0.45	0.43	0.54
	1GU4	0.38	0.34	1.36
	1GU5	2.05	1.22	1.50
	1GU6	0.84	0.34	0.74
2GU	2GU1	1.52	0.44	1.15
	2GU2	0.65	0.84	0.78
	2GU3	0.38	0.52	0.48
	2GU4	0.45	0.56	0.39
	2GU5	1.10	0.35	1.05
	2GU6	2.20	0.42	2.37
3GU	3GU1	2.24	0.63	2.34
	3GU2	0.48	0.63	0.51
	3GU3	0.43	0.67	0.51
	3GU4	1.65	0.68	1.60
	3GU5	1.58	0.50	1.21
	3GU6	0.97	0.68	0.78
4GU	4GU1	0.51	0.37	0.78
	4GU2	1.40	0.55	1.17
	4GU3	0.39	0.73	0.46
	4GU4	0.69	0.83	0.69
	4GU5	0.74	0.73	0.78

Table A.6. Range of Magnitudes of S_A^2, S_B^2, and S_{KS100}^2

RUN SERIES	FIT	S_A^2			S_B^2			S_{KS100}^2		
1LN	ML	0.00086	TO	0.00130	0.00033	TO	0.00055	0.0074	TO	0.0223
	LS	0.00098	TO	0.00143	0.00081	TO	0.00125	0.0159	TO	0.0331
2LN	ML	0.00102	TO	0.00193	0.00053	TO	0.00102	0.0146	TO	0.0392
	LS	0.00151	TO	0.00228	0.00120	TO	0.00209	0.0506	TO	0.0636
3LN	ML	0.00160	TO	0.00240	0.00081	TO	0.00127	0.0511	TO	0.1444
	LS	0.00239	TO	0.00417	0.00224	TO	0.00327	0.1283	TO	0.1452
4LN	ML	0.00313	TO	0.00340	0.00152	TO	0.00312	0.1232	TO	0.4058
	LS	0.00454	TO	0.00702	0.00389	TO	0.00745	0.5432	TO	0.8046
1GA	ML	0.21	TO	0.35	0.18	TO	0.26	0.0424	TO	0.0510
	LS	0.40	TO	0.96	0.31	TO	1.03	0.0841	TO	0.1452
2GA	ML	0.54	TO	0.71	0.41	TO	0.64	0.0151	TO	0.0353
	LS	0.82	TO	1.37	0.76	TO	1.19	0.0280	TO	0.0506
3GA	ML	0.46	TO	1.26	0.39	TO	0.83	0.0062	TO	0.0253
	LS	0.76	TO	2.57	0.94	TO	1.91	0.0064	TO	0.0437
4GA	ML	1.10	TO	2.00	0.97	TO	1.35	0.0069	TO	0.0100
	LS	3.05	TO	3.50	2.45	TO	3.26	0.0125	TO	0.0237
1GU	MO	0.0174	TO	0.0403	0.0022	TO	0.0056	0.0676	TO	0.1444
	LS	0.0141	TO	0.0460	0.0035	TO	0.0177	0.0697	TO	0.1347
2GU	MO	0.0323	TO	0.0639	0.0021	TO	0.0033	0.0416	TO	0.0835
	LS	0.0291	TO	0.1004	0.0037	TO	0.0073	0.0353	TO	0.1755
3GU	MO	0.0459	TO	0.1329	0.0010	TO	0.0017	0.0154	TO	0.0372
	LS	0.0572	TO	0.1500	0.0019	TO	0.0027	0.0154	TO	0.0320
4GU	MO	0.1144	TO	0.1786	0.0005	TO	0.0008	0.0092	TO	0.0149
	LS	0.1143	TO	0.2933	0.0006	TO	0.0019	0.0125	TO	0.0210

ϕ=1.0 (see Equation A.20) the weighted least squares method reduces to the method of minimum chi-square which is known to be asymptotically equivalent to the method of maximum likelihood (Kendall and Stuart, 1973). With the use of non-zero values of ϕ it may be expected that some of the characteristics of the unweighted least squares method, such as relatively low efficiency may be corrected. In order to determine the effect of weight on different properties of least squares estimates some of the runs of each series of runs, shown in Table A.1 were repeated with values assigned to ϕ in the range of 0 to 1.0. The results of weighted least squares are summarized in Tables A.7 through A.9 for some of the runs of lognormal, gamma and Gumbel distributions. These tables show the mean values and variances of parameter estimates and the 100-year predictions for each run. In addition, the ratios of mean 100-year predictions ($\bar{K}_{S100}$), and the corresponding population prediction ($K_{100}=K_{.99}$) are also shown.

The effects of weights on the least squares estimates are as follows:

The Effect of Weight on Percentiles The overall effect of weights on the fit is such that the predictions (for return period 10 years and above) have invariably increased for the three distributions studied. The larger the value of ϕ, the larger is the increase in percentiles compared to the LS percentiles with ϕ=0.00. This effect of increase in percentiles with the use of weight helps to eliminate the negative bias in LS percentiles observed with gamma distribution and with certain runs of lognormal and Gumbel distributions when

Table A.7. Results of Weighted LS - LN PDF

(n = 100)

RUN NO	FIT	ϕ	$\bar{A}$	S_A^2	$\bar{B}$	S_B^2	$\bar{K}_{S100}$	S^2_{KS100}	$\frac{\bar{K}_{S100}}{K_{100}}$
1LN2	ML		-0.041	0.00085	0.300	0.00035	1.931	0.0094	1.0054
	LS	0.00	-0.036	0.00139	0.304	0.00101	1.954	0.0289	1.0225
	LS	0.25	-0.035	0.00134	0.305	0.00092	1.971	0.0279	1.0260
	LS	0.50	-0.033	0.00130	0.307	0.00081	1.990	0.0269	1.0308
	LS	0.75	-0.032	0.00128	0.309	0.00070	1.994	0.0255	1.0379
1LN5	ML		-0.045	0.00130	0.293	0.00055	1.916	0.0223	0.9974
	LS	0.00	-0.042	0.00162	0.300	0.00085	1.933	0.0328	1.0061
	LS	1.00	-0.038	0.00116	0.307	0.00054	1.972	0.0199	1.0268
2LN2	ML		-0.086	0.00183	0.403	0.00070	2.341	0.0286	1.0015
	LS	0.00	-0.090	0.00229	0.403	0.00159	2.345	0.0543	1.0018
	LS	0.25	-0.089	0.00210	0.404	0.00141	2.353	0.0502	1.0054
	LS	0.50	-0.087	0.00195	0.406	0.00120	2.368	0.0449	1.0117
	LS	0.75	-0.083	0.00189	0.410	0.00096	2.395	0.0384	1.0239
2LN5	ML		-0.071	0.00102	0.404	0.00102	2.392	0.0392	1.0220
	LS	0.00	-0.076	0.00151	0.394	0.00150	2.333	0.0586	0.9969
	LS	1.00	-0.065	0.00104	0.414	0.00126	2.467	0.0506	1.0542
3LN3	ML		-0.132	0.00240	0.501	0.00170	2.831	0.1444	1.0027
	LS	0.00	-0.147	0.00341	0.495	0.00231	2.753	0.1452	0.9748
	LS	0.25	-0.144	0.00272	0.498	0.00225	2.782	0.1354	0.9851
	LS	0.50	-0.139	0.00227	0.502	0.00215	2.820	0.1267	0.9987
	LS	0.75	-0.131	0.00218	0.508	0.00200	2.879	0.1260	1.0193
3LN4	ML		-0.114	0.00160	0.505	0.00159	2.905	0.0888	1.0287
	LS	0.00	-0.116	0.00239	0.491	0.00327	2.816	0.1544	0.9975
	LS	1.00	-0.106	0.00176	0.521	0.00229	3.043	0.1347	1.0776
4LN2	ML		-0.240	0.00424	0.696	0.00152	3.934	0.1232	0.9990
	LS	0.00	-0.242	0.00702	0.688	0.00425	3.953	0.5432	0.9912
	LS	0.25	-0.239	0.00652	0.691	0.00388	3.935	0.4970	0.9994
	LS	0.50	-0.235	0.00580	0.698	0.00334	4.053	0.4277	1.0153
	LS	0.75	-0.231	0.00500	0.710	0.00272	4.179	0.3505	1.0481
4LN4	ML		-0.258	0.00403	0.687	0.00265	3.956	0.2841	0.9571
	LS	0.00	-0.263	0.00698	0.687	0.00542	3.881	0.7157	0.9734
	LS	1.00	-0.242	0.00586	0.704	0.00241	4.074	0.2735	1.0215

Table A.8. Results of Weighted LS - GA PDF

(n = 100)

RUN NO	FIT	ϕ	$\bar{A}$	S_A^2	$\bar{B}$	S_B^2	$\bar{K}_{S100}$	S^2_{KS100}	$\bar{K}_{S100}/K_{100}$
1GA1	ML		3.00	0.21	3.02	0.18	2.845	0.0610	1.0153
	LS	0.00	3.20	0.48	3.19	0.32	2.765	0.1399	0.9869
	LS	0.50	3.09	0.37	3.07	0.26	2.812	0.1156	1.0036
	LS	0.75	2.97	0.30	2.99	0.22	2.872	0.1024	1.0250
1GA2	ML		3.05	0.35	3.09	0.19	2.827	0.0605	1.0091
	LS	0.00	3.24	0.40	3.22	0.35	2.755	0.0957	0.9834
	LS	0.50	3.14	0.29	3.14	0.25	2.795	0.0801	0.9977
	LS	0.75	3.04	0.24	3.07	0.20	2.842	0.0724	1.0142
1GA3	ML		3.22	0.24	3.20	0.23	2.740	0.0424	0.9766
	LS	0.00	3.47	0.96	3.44	1.03	2.690	0.0930	0.9600
	LS	1.00	3.04	0.27	3.07	0.27	2.837	0.0595	1.0125
1GA5	ML		3.03	0.29	2.95	0.23	2.780	0.0686	0.9931
	LS	0.00	3.05	0.44	2.96	0.31	2.805	0.1452	1.0000
	LS	1.00	2.84	0.40	2.83	0.28	2.930	0.1444	1.0458
2GA1	ML		5.17	0.54	5.14	0.52	2.300	0.0185	0.9911
	LS	0.00	5.33	0.90	5.28	0.90	2.277	0.0339	0.9811
	LS	0.75	5.01	0.51	5.02	0.58	2.337	0.0234	1.0071
2GA2	ML		5.25	0.65	5.15	0.53	2.273	0.0253	0.9792
	LS	0.00	5.59	0.82	5.42	0.76	2.209	0.0289	0.9517
	LS	1.00	4.99	0.70	4.97	0.54	2.341	0.0350	1.0086
2GA5	ML		5.05	0.56	4.98	0.64	2.300	0.0151	0.9908
	LS	0.00	5.08	1.37	5.01	1.19	2.326	0.0506	1.0023
	LS	1.00	4.61	0.52	4.63	0.53	2.414	0.0310	1.0402
3GA1	ML		7.13	0.58	7.11	0.77	2.072	0.0062	0.9955
	LS	0.00	7.64	0.76	7.69	0.96	2.040	0.0064	0.9803
	LS	0.75	7.11	0.85	7.14	0.91	2.087	0.0038	1.0025

Table A.8. -Continued

RUN NO	FIT	ϕ	$\bar{A}$	S_A^2	$\bar{B}$	S_B^2	$\bar{K}_{S100}$	S^2_{KS100}	$\bar{K}_{S100}/K_{100}$
3GA2	ML		7.08	1.25	7.02	0.88	2.091	0.0253	0.9998
	LS	0.00	7.12	2.57	7.00	1.91	2.085	0.0437	1.0016
	LS	0.75	6.88	1.39	6.85	0.98	2.109	0.0310	1.0133
3GA3	ML		7.10	0.46	7.05	0.39	2.070	0.0090	0.9943
	LS	0.00	7.12	0.91	7.13	0.94	2.083	0.0159	1.0009
	LS	0.75	6.93	0.66	6.93	0.53	2.099	0.0154	1.0085
3GA4	ML		7.26	0.68	7.23	0.50	2.071	0.0125	0.9951
	LS	0.00	7.39	1.69	7.34	1.32	2.062	0.0286	0.9907
	LS	1.00	6.85	0.83	6.85	0.72	2.129	0.0174	1.0229
3GA5	ML		6.98	0.84	6.90	0.68	2.076	0.0139	0.9973
	LS	0.00	7.02	1.81	6.94	1.24	2.092	0.0365	1.0052
	LS	1.00	6.55	1.20	6.55	0.88	2.149	0.0276	1.0325
4GA1	ML		10.16	1.47	10.04	1.11	1.862	0.0100	0.9912
	LS	0.00	10.75	3.05	10.51	2.45	1.839	0.0237	0.9744
	LS	0.75	10.14	1.88	10.01	1.49	1.866	0.0149	0.9937
4GA2	ML		10.29	1.97	10.31	1.66	1.876	0.0090	0.9937
	LS	0.00	10.38	3.16	10.48	3.06	1.887	0.0139	1.0046
	LS	0.75	10.01	2.95	10.09	2.38	1.905	0.0145	1.0141
4GA4	ML		10.13	2.00	10.07	1.77	1.871	0.0093	0.9953
	LS	0.00	10.51	3.30	10.48	2.93	1.868	0.0210	0.9943
	LS	1.00	9.69	2.31	9.72	2.19	1.911	0.0121	1.0173
4GA5	ML		9.98	1.10	9.91	0.97	1.875	0.0059	0.9983
	LS	0.00	10.48	3.08	10.48	2.71	1.867	0.0125	0.9941
	LS	1.00	9.57	1.05	9.58	0.87	1.908	0.0075	1.0158
4GA3	ML		10.53	1.87	10.44	1.85	1.847	0.0071	0.9832
	LS	0.00	10.36	3.50	10.33	3.25	1.879	0.0144	1.0040
	LS	0.75	10.16	2.42	10.17	2.21	1.893	0.0108	1.0024

Table A.9. Results of Weighted LS - GU PDF

(n = 100)

RUN NO	FIT	ϕ	$\bar{A}$	S_A^2	$\bar{B}$	S_B^2	$\bar{K}_{S100}$	S^2_{KS100}	$\frac{\bar{K}_{S100}}{K_{100}}$
1GU1	MO		1.627	0.0382	0.7361	0.0022	3.605	0.1362	0.3732
	LS	0.00	1.682	0.0163	0.6159	0.0107	3.363	0.0697	0.9147
	LS	1.00	1.499	0.0163	0.6385	0.0052	3.729	0.0560	1.0127
1GU5	MO		1.587	0.0289	0.7777	0.0041	3.710	0.1095	1.0077
	LS	0.00	1.581	0.0141	0.6718	0.0036	3.593	0.0729	0.9772
	LS	1.00	1.460	0.0109	0.6762	0.0035	3.843	0.0489	1.0437
1GU6	MO		1.588	0.0240	0.7584	0.0056	3.683	0.0992	1.0002
	LS	0.00	1.680	0.0287	0.6517	0.0177	3.419	0.1347	0.9284
	LS	1.00	1.510	0.0168	0.6699	0.0099	3.739	0.0778	1.0154
2GU1	MO		2.083	0.0375	0.7294	0.0026	2.956	0.0433	0.9817
	LS	0.00	2.034	0.0247	0.6956	0.0059	2.971	0.0376	0.9867
	LS	1.00	1.940	0.0176	0.7035	0.0039	3.091	0.0292	1.0264
2GU5	MO		2.048	0.0457	0.7592	0.0021	3.030	0.0610	1.0062
	LS	0.00	2.005	0.0417	0.7290	0.0060	3.045	0.0581	1.0112
	LS	1.00	1.903	0.0207	0.7385	0.0034	3.170	0.0345	1.0526
2GU6	MO		2.131	0.0639	0.7278	0.0031	2.918	0.0835	0.9689
	LS	0.00	2.100	0.0291	0.6955	0.0073	2.901	0.0353	0.9632
	LS	1.00	1.958	0.0239	0.7067	0.0051	3.059	0.0342	1.0156
3GU1	MO		3.134	0.1329	0.8047	0.0015	2.293	0.0361	0.9795
	LS	0.00	3.138	0.0592	0.7948	0.0024	2.269	0.0154	0.9695
	LS	1.00	3.008	0.0487	0.8114	0.0017	2.349	0.0159	1.0034
3GU5	MO		2.995	0.0902	0.8193	0.0010	2.370	0.0262	1.0125
	LS	0.00	2.952	0.0572	0.8166	0.0020	2.385	0.0216	1.0189
	LS	1.00	2.857	0.0436	0.8233	0.0010	2.442	0.0154	1.0430
3GU6	MO		3.062	0.0816	0.8052	0.0015	2.320	0.0198	0.9909
	LS	0.00	3.022	0.0837	0.7996	0.0022	2.336	0.0240	0.9978
	LS	1.00	2.954	0.0558	0.8119	0.0016	2.379	0.0156	1.0161
4GU1	MO		4.009	0.1273	0.8625	0.0007	2.020	0.0145	1.0053
	LS	0.00	4.171	0.2494	0.8533	0.0019	1.989	0.0198	0.9856
	LS	1.00	3.911	0.1070	0.8661	0.0003	2.059	0.0132	1.0216
4GU5	MO		4.044	0.1470	0.8561	0.0003	2.004	0.0149	0.9983
	LS	0.00	4.002	0.1999	0.8532	0.0011	1.991	0.0190	0.9919
	LS	1.00	3.957	0.1221	0.8607	0.0009	2.032	0.0149	1.0125

unitary weights are used. While a ϕ value of 0.25 does not have any appreciable effect on the results (see Table A.7), values of 0.75 and 1.00 have virtually eliminated all the negative bias in percentiles wherever observed. Use of ϕ larger than 1.0 is not warranted since it would result in larger positive bias of the predictions. Use of ϕ=1.00 brings the fit for Gumbel distributions close to the population distributions. The causes for increase in the value of percentiles with the use of weight may be explained as follows:

The sum of the squares errors which is sought to be minimized in the least squares method may be given in its general form by

$$SSE = \sum_{i=1}^{N} (h_i - \bar{p}_i)^2/(\bar{p}_i)^{\phi} \qquad A.47$$

By equation A.47 for non-zero values of ϕ the error terms occurring away from the mode of the distribution are given a larger weight since $\bar{p}_i$ becomes an extremely small value in the tail of the distribution. (It may be noted that the $\bar{p}_i$ is always less than unity. A typical low value of $\bar{p}$ observed for the extreme right tail class interval was 0.0002). Thus the weighted least squares method attempts to use as much of tail information as possible, i.e., it attempts to fit the tail better, particularly the right tail (which extends to infinity). In general, the values of h_i will be much larger than $\bar{p}_i$ for right tail class intervals which results in a

thicker right tail for weighted LS fits compared to the right tail of unweighted LS fit (see Figure A.3). Such a thickening of tail takes place at the expense of some area under density curve being shifted to the right which is responsible for the increase in the values of percentiles. The opposite effect also may occur for samples which do not have any observations falling the class intervals of thinner right tail but have observations in the left tail. In this case the left tail is thickened shifting some area of the density curve to the left and the k values for different higher CDF's will be lowered. On the basis of above analysis it may be stated that when data samples contain items in the thinner tail portions of a proposed PDF weighted (ϕ=0.75 or 1.00) LS method will attempt to fit the tails better than the LS method with ϕ=0.00 (see also section "Samples with Outliers", Chapter VI).

The Effect of Weight on Parameters With the use of weights, the mean estimates of both the lognormal parameters increased, both the gamma parameters decreased and for Gumbel distribution parameter A increased while the parameter B decreased. (See Tables A.7 through A.9). The effect is, in general, to increase the value of k for a given return period. The reason for such a happening is pictorially explained by Figure A.3.

The Effect of Weight on the Efficiency of Parameter Estimates Use of weights in general decreases the variance of parameter estimates, S_A^2 and S_B^2, and in most cases the values of least squares

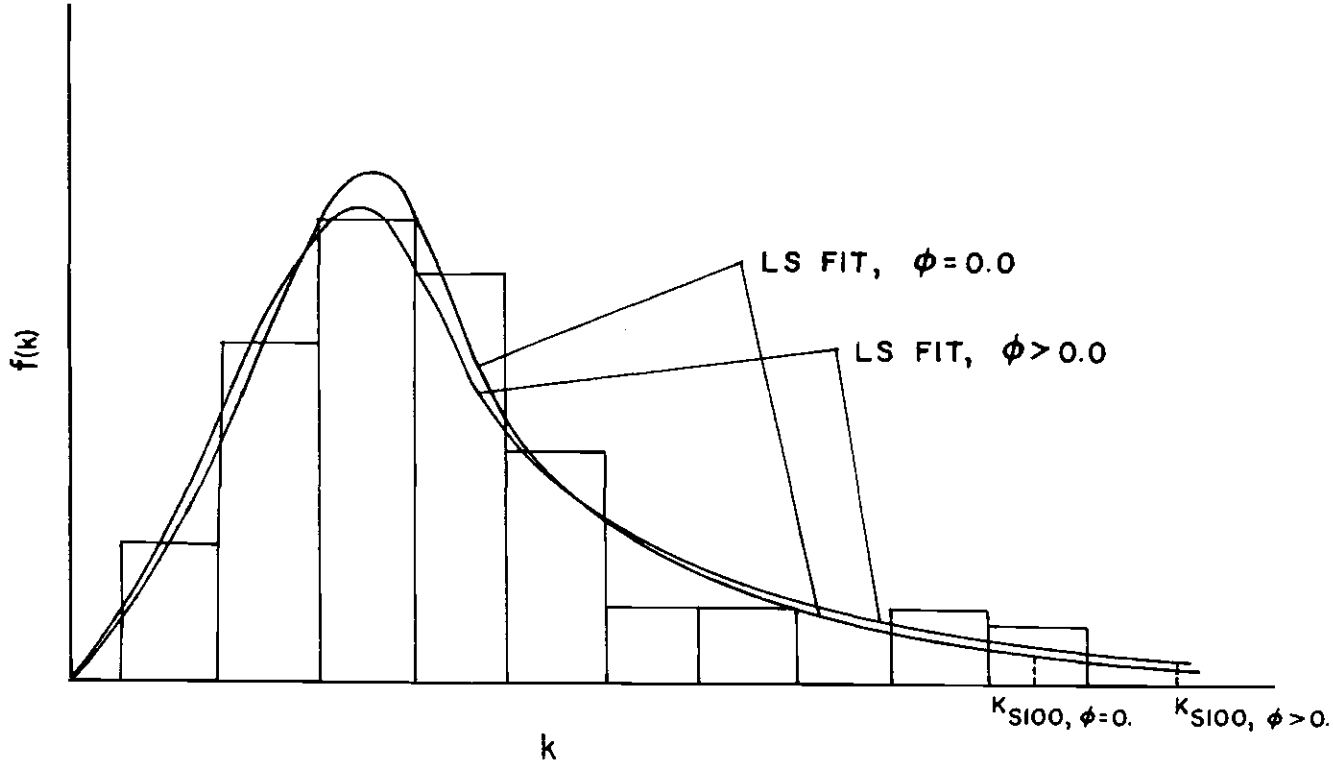


Figure A.3 The Effect of Weight on Least Squares Fit

variances approach those of maximum likelihood when ϕ=1.00 for lognormal and gamma distributions (see Tables A.7 and A.8). Such an improvement in the variance of least squares estimates is not wholly unexpected. Since in the case ϕ=1.00 the method of weighted lease squares is equivalent to the method of minimum chi-square, and the ML and MCS methods have the same asymptotic properties (see Kendall and Stuart, 1973). For other values of ϕ the LS estimates may be expected to possess the properties of ML in varying degrees. In case of Gumbel distribution the least squares method with ϕ=1.00, perhaps, offers an alternative to maximum likelihood since ML equations are difficult to solve for this distribution. The procedure described by Harter and Moore (1967) to obtain maximum likelihood estimates of Gumbel distribution, assumes that the scale parameter α is known. They obtain an explicit expression for the ML estimator of location parameter u only by assuming that the scale parameter α is known. Beard (1974) uses this method iteratively to obtain ML estimates of Gumbel distribution in his recent work on Flood/Flow Frequency Techniques.

Table A.9 shows that the sample variance of LS 100-year predictions with ϕ=1.00 is invariably smaller than the variance based on the moments method for Gumbel distribution. This establishes the superiority of the (weighted) lease squares method over the method of

moments for Gumbel distributions with respect to efficiency.

Summary Use of weight reduces bias and improves efficiency of least squares estimates. Use of ϕ in the range of 0.75 to 1.00 eliminated the negative bias of predictions (i.e., percentiles) with gamma distribution and with some runs of lognormal and Gumbel distributions. In case of Gumbel distribution, the least squares method with ϕ=1.0 offers an alternative to maximum likelihood method.

Study 4: The Effect of a "Growing" Sample upon Frequency Analysis.

One feature of the hydrologic data is the addition of new data as time passes. This may be viewed as "growth" of a sample since a hydrologic data sample literally grows with time. One of the concerns of a hydrologist conducting a frequency analysis is whether the results of the analysis made with the available data will be greatly affected by data that would be available a few years in the future.

For a given hydrologic station, it may be reasonably expected that the causative factors leading to the occurrence of various hydrologic events are constant with time. Statistically this is equivalent to saying that the data at a given hydrologic station are stationary. Then, except for the vagaries of the sample size a "growing" sample should not affect a frequency analysis in any way. To examine this aspect of frequency analysis, particularly the working of the least squares method, the computer was programmed to "grow" a sample in which the larger samples retained the data of smaller samples (see Appendix E for details).

For each of the three probability distributions simulations runs were made with sample sizes 25, 50, 75 and 100. These runs were made keeping the population parameters constant for a set of four runs having the above mentioned sample sizes. Since a hydrologists main concern is with the sample predictions, the statistical properties of the percentiles (100-year, i.e. $K_{.99}$, event is typically chosen) rather than the parameters of the distribution are examined as the sample size varied. Tables A.10 through A.12 summarize the results of growing samples for LN, GA and GU distributions, respectively. These tables show for each run, the sample mean and the sample standard deviation of the 100-year predictions, and the ratio of the sample mean and the population value of 100-year predictions as given by the least squares and the maximum likelihood (moments in case of Gumbel distribution) methods.

Tables A.10 through A.12 show that the LS mean sample predictions, in general, are stable with sample size (the lowest sample size = 25) for all the three distributions. The error in predictions is in the range of -4% to +7% for samples of 25 size and becomes much less for the samples of larger size. For LN and GA distributions, in general, neither LS nor ML mean predictions showed any superiority over each other except that the cases of high positive error (+7%) did not occur in the ML predictions of 25 size samples. In case of Gumbel distribution also neither LS nor the MO method may be stated superior to one another on the basis of mean sample predictions.

The standard deviations of predictions from 25 size samples are,

Table A.10. Growing Sample Analysis - LN PDF

The Statistical Characteristics of the 100 - Year Sample Predictions

(ϕ = 0.00)

σ_K^2	SAMPLE SIZE	--- LS --- $\overline{K}_S$	$S(K_S)$	$\overline{K}_S/K_P$	--- MO --- $\overline{K}_S$	$S(K_S)$	$\overline{K}_S/K_P$	RUN NO
0.7311	25	3.726	0.837	1.0120	3.874	0.626	1.0520	1GU5
	50	3.648	0.457	0.9909	3.729	0.494	1.0129	
	75	3.570	0.327	0.9697	3.691	0.395	1.0025	
	100	3.593	0.270	0.9772	3.710	0.331	1.0077	
0.4112	25	3.199	0.592	1.0521	3.147	0.384	1.0450	2GU5
	50	3.124	0.322	1.0373	3.078	0.351	1.0220	
	75	3.041	0.262	1.0098	3.028	0.298	1.0055	
	100	3.045	0.241	1.0112	3.030	0.247	1.0062	
0.1829	25	2.443	0.416	1.0435	2.450	0.245	1.0465	3GU5
	50	2.424	0.248	1.0353	2.395	0.229	1.0232	
	75	2.385	0.161	1.0190	2.365	0.194	1.0101	
	100	2.385	0.147	1.0189	2.370	0.162	1.0126	
0.1208	25	1.959	0.259	0.9812	2.044	0.218	1.0183	4GU5
	50	1.968	0.175	0.9807	2.009	0.151	1.0010	
	75	1.998	0.162	0.9955	2.011	0.143	1.0022	
	100	1.991	0.138	0.9919	2.004	0.122	0.9983	

Table A.11. Growing Sample Analysis - GA PDF

The Statistical Characteristics of the 100 - Year Sample Predictions

(ϕ = 0.00)

σ_K^2	SAMPLE SIZE	--- LS --- $\bar{K}_S$	$S(K_S)$	$\bar{K}_S/K_P$	--- ML --- $\bar{K}_S$	$S(K_S)$	$\bar{K}_S/K_P$	RUN NO
0.3333	25	2.836	0.802	1.0123	2.750	0.528	0.9820	1GA5
	50	2.780	0.512	0.9922	2.779	0.370	0.9919	
	75	2.857	0.470	1.0196	2.795	0.314	0.9976	
	100	2.805	0.381	1.0000	2.780	0.262	0.9931	
0.2000	25	2.249	0.442	0.9590	2.228	0.290	0.9597	2GA5
	50	2.300	0.339	0.9903	2.291	0.149	0.9873	
	75	2.318	0.271	0.9989	2.292	0.150	0.9874	
	100	2.326	0.225	1.0023	2.300	0.123	0.9908	
0.1429	25*	2.222	0.307	1.0577	2.077	0.242	0.9977	3GA5
	50@	2.145	0.314	1.0306	2.075	0.194	0.9969	
	75	2.084	0.216	1.0010	2.054	0.139	0.9869	
	100	2.092	0.191	1.0052	2.076	0.118	0.9973	
0.1000	25#	1.892	0.209	1.0072	1.845	0.181	0.9821	4GA5
	50@	1.875	0.135	0.9984	1.874	0.121	0.9976	
	75	1.879	0.129	1.0004	1.879	0.091	1.0004	
	100	1.867	0.112	0.9941	1.875	0.083	0.9983	

* -CONVERGENCE = 78%
@ -CONVERGENCE = 96%
\# -CONVERGENCE = 92%

Table A.12. Growing Sample Analysis - GU PDF

The Statistical Characteristics of the 100 - Year Sample Predictions

(ϕ = 0.00)

σ_K^2	SAMPLE SIZE	--- LS --- $\bar{K}_S$	$S(K_S)$	$\bar{K}_S/K_P$	--- ML --- $\bar{K}_S$	$S(K_S)$	$\bar{K}_S/K_P$	RUN NO
0.0942	25	1.839	0.302	0.9575	1.929	0.256	1.0041	1LN5
	50	1.876	0.238	0.9766	1.900	0.217	0.9392	
	75	1.905	0.204	0.9915	1.912	0.181	0.9956	
	100	1.933	0.181	1.0061	1.916	0.151	0.9974	
0.1735	25	2.356	0.436	1.0066	2.327	0.377	0.9943	2LN2
	50	2.422	0.386	1.0349	2.391	0.246	1.0216	
	75	2.386	0.304	1.0193	2.366	0.202	1.0108	
	100	2.345	0.233	1.0018	2.344	0.169	1.0015	
0.2840	25	2.768	1.159	0.9805	2.750	0.554	0.9741	3LN1
	50	2.860	0.632	1.0130	2.804	0.351	0.9932	
	75	2.890	0.585	1.0235	2.796	0.382	0.9902	
	100	2.804	0.474	0.9931	2.789	0.273	0.9878	
0.6323	25	4.262	2.136	1.0689	3.926	1.242	0.9845	4LN4
	50	3.828	0.958	0.9600	3.810	0.656	0.9555	
	75	3.898	0.834	0.9776	3.778	0.465	0.9474	
	100	3.881	0.846	0.9734	3.856	0.533	0.9671	

on an average, about twice the standard deviations of the predictions from samples of size 100 for all the methods. Statistically, the results obtained in this study may be interpeted as follows: There does not appear to be any particular advantage in any one method of parameter estimation when viewed in the context of growing samples (see Chapter VI for results of frequency analysis with "growing" samples of real data).

Study 5: Distribution of Errors in Least Square Fittings

If the error terms e_i in Equation A.1 are normally distributed with zero mean and constant (unknown) variance, σ^2, the principles of LS and ML are equivalent (Graybill (1961), Grant (1973)). However, since the observations were grouped in applying LS method in this work (see Equation A.18) the solutions given by LS and ML may not be equivalent even if e_i are normally distributed. But if e_i in Equation A.18 are $N(0,\sigma^2)$, the LS method examined in this work may be broadly regarded as an application of the concepts of ML. To examine empirically the distribution of errors resulting from the least squares method, the chi-square and the Kolmogorov-Smirnov (K-S) goodness-of-fit tests for normality were performed on the error terms arising in the fitting of each sample (see Chapter IV for a description of these tests). In these tests, the population mean (of errors) has been assumed to be zero and the sample variance of errors is treated as the population variance of errors. For the chi-square test, the data were so grouped that expected number of occurrence of errors in each group would be at least three.

To test the hypothesis that the residual least square errors are normally distributed with zero means the simulated data samples were divided into the following categories:

1. Data samples fit to the parent distribution, sample size = 100, and ϕ=0.00.
2. Data samples fit to the parent distribution, sample size less than 100 (samples of size 25, 50 and 75 in equal number are included in this category) and ϕ=0.00.
3. Data samples fit to the parent distribution. Errors are weighted and ϕ=0.75 or 1.00.
4. Data samples fit to other than parent distribution and ϕ=0.00.

The aim of using the above classification of data samples is to examine whether the "normality-of-errors" hypothesis is more closely valid for certain categories of data. In addition, these tests were performed for various levels of σ_k^2 for each of the three distributions.

The Results of the Chi-square test Tables A.13 through A.15 summarize the results for chi-square test for LN, GA and GU distributions, respectively. It may be recalled that the statistic δ defined by

$$P(\chi^2 \geq \chi_o^2) = \delta$$

in which χ_o^2 is given by Equation 4.1, is a measure of closeness of fit. A large value of δ indicates that the set of errors being tested is in fact distributed approximately in a normal fashion while the hypothesis that

the errors are normally distributed with zero mean may be rejected if $\delta \leq \alpha$, where α is the probability of rejecting the null hypothesis when it is true, that is, α=P(reject H_o/H_o is true) . This investigation yielded values of δ ranging from near zero to near unity which shows that the hypothesis may be accepted for some groups of errors and rejected for others.

Based on the distribution of δ (see Tables A.13 through A.15) with σ_k^2 for different categories of simulated samples, the following observations may be made:

1. The number of rejections of null hypothesis at 5% significance level is more at low variances of the data samples ($\sigma_k^2 \approx 0.1$) and falls sharply as σ_k^2 increases. The percentage rejections is about 20 to 32 when $\sigma_k^2 \approx 0.1$ and 5.5 to 17 when σ_k^2 is larger (0.14 and above) the higher figures being for lognormal distribution.
2. The sample size of the data and the fact that the data of one distribution are fit to another distribution do not affect the results much. The percent rejections, at a given level of σ_k^2, for these categories is approximately same as that for category 1.
3. In general, the percent rejections of null hypothesis is reduced for the error terms of weighted least squares fits with ϕ=0.75 or 1.00 compared to the same for the error terms with ϕ=0.00.

The above observations based on chi-square test indicate that in a large majority of cases and under a large variety of data conditions the

Table A.13. Distribution of δ for Error Terms - LN PDF

σ_K^2	CATEGORY*	NO OF SAMPLES EXAMINED	NUMBER OF OCCURRENCES $\delta = P(X^2 \geq X_o^2)$ .0-.05	.05-.1	.1-.2	.2-.4	.4-.6	.6-.8	.8-1.0
0.0942	1	125	40	15	20	27	11	6	6
	2	75	27	13	15	7	9	2	2
	3	75	17	8	9	18	12	5	6
	4	49	19	5	3	7	10	2	3
	SUB-TOTAL	324	103	41	47	59	42	15	17
0.1735	1	125	18	15	20	20	21	17	14
	2	75	5	7	14	15	14	9	11
	3	75	9	11	5	11	7	15	17
	4	49	12	6	3	11	8	6	3
	SUB-TOTAL	324	44	39	42	57	50	47	45
0.2840	1	100	21	7	11	18	16	14	13
	2	75	8	5	8	14	22	9	9
	3	50	7	2	6	4	8	14	9
	4	50	10	3	13	6	7	9	2
	SUB-TOTAL	275	46	17	38	42	53	46	33
0.6323	1	100	17	4	15	20	19	15	10
	2	75	9	4	8	15	18	6	15
	3	75	17	6	10	14	9	12	7
	4	47	9	2	6	10	8	7	5
	SUB-TOTAL	297	52	16	39	59	54	40	37

* -SEE CATEGORY DEFINITION, PAGE

Table A.14. Distribution of δ for Error Terms - GA PDF

σ_K^2	CATEGORY*	NO OF SAMPLES EXAMINED	NUMBER OF OCCURRENCES $\delta = P(X^2 \geq X_o^2)$ .0-.05	.05-.1	.1-.2	.2-.4	.4-.6	.6-.8	.8-1.0
0.1000	1	121	31	25	18	14	20	8	5
	2	71	16	10	9	12	11	8	5
	3	125	16	13	12	31	20	12	21
	4	50	13	12	3	9	3	3	3
	SUB-TOTAL	367	75	60	47	65	54	31	34
0.1429	1	125	12	11	12	33	25	20	12
	2	67	7	3	10	11	15	7	14
	3	125	10	3	14	29	22	21	26
	4	50	8	7	3	9	9	7	7
	SUB-TOTAL	367	37	24	39	82	71	55	59
0.2000	1	125	13	6	15	25	25	23	19
	2	73	5	6	7	17	15	9	13
	3	75	4	2	5	15	21	13	15
	4	50	6	3	3	8	16	10	4
	SUB-TOTAL	323	29	17	30	63	77	55	51
0.3333	1	125	7	2	24	28	21	24	19
	2	75	2	4	8	22	20	7	12
	3	100	11	6	18	14	18	16	17
	4	50	5	3	7	9	13	5	8
	SUB-TOTAL	350	25	15	57	73	72	32	56

* -SEE CATEGORY DEFINITION, PAGE

Table A.15. Distribution of δ for Error Terms - GU PDF

σ_K^2	CATEGORY*	NO OF SAMPLES EXAMINED	NUMBER OF OCCURRENCES $\delta = P(\chi^2 \geq \chi_o^2)$ .0-.05	.05-.1	.1-.2	.2-.4	.4-.6	.6-.8	.8-1.0
0.1028	1	125	32	19	23	16	17	10	8
	2	75	19	11	13	10	11	8	3
	3	50	7	6	4	9	12	7	5
	4	48	12	7	9	9	7	1	3
	SUB-TOTAL	298	70	43	49	44	47	26	19
0.1828	1	125	18	3	24	21	26	15	18
	2	75	4	3	10	13	15	13	17
	3	75	9	1	11	11	15	19	9
	4	50	6	3	12	8	12	7	2
	SUB-TOTAL	325	37	10	57	53	68	54	46
0.4112	1	150	6	9	17	37	27	26	28
	2	75	4	5	11	13	20	5	17
	3	75	6	5	6	13	23	8	14
	4	50	3	2	3	11	11	9	11
	SUB-TOTAL	350	19	21	37	74	81	48	70
0.7311	1	150	14	16	10	41	38	11	20
	2	75	5	7	12	17	15	15	4
	3	75	4	7	8	26	22	3	5
	4	49	4	3	2	13	12	6	9
	SUB-TOTAL	349	27	33	32	97	87	35	38

* -SEE CATEGORY DEFINITION, PAGE

hypothesis that the error terms of least squares fit is normally distributed with zero mean is not rejected at 5% significance level. On the whole the chi-square test would require rejection of 567 cases in 3949 cases which is about 14% of the total. However, it may be noted that the chi-square goodness-of-fit test is actually a large sample test and is not wholly appropriate for application in this instance since the sets of errors tested are not large sets (the number of errors typically ranged from 16 to 20). On the other hand, the Kolmogorov-Smirnov goodness-of-fit test is an exact test for all sample sizes (Benjamin and Cornell, 1970). The results of Kolmogorov-Smirnov test are discussed below.

The Results of Kolmogorov-Smirnov Test The test statistic, D_o of K-S goodness-of-fit test is computed (by the procedure described in Chapter IV) assuming that the error terms of least squares fit are distributed normally with zero mean and variance equal to the sample variance of the error terms. The hypothesis that the errors are distributed normally with mean zero is rejected if D_o is larger than a critical value at a given significance levels. All the samples examined under chi-square test are also subjected to K-S test for normality of errors and the number rejections of these samples against the null hypothesis has been evaluated at significant levels α=0.1 and 0.05. The number of rejections of the null hypothesis by K-S test at α=5% is found to be far below than the same by chi-square test. Hence, no attempt has been made to present K-S test results in detail for each category of data samples mentioned earlier. Table A.16 summarizes the results of K-S normality

Table A.16. Results of Kolmogorov - Smirnov Test for Normality of Errors

Distribution	σ_K^2	No of Samples Examined	Number of rejections $\alpha = 0.10$	$\alpha = 0.05$
Lognormal	0.0942	324	8	4
	0.1735	324	4	2
	0.2840	275	5	2
	0.6323	297	11	3
Gamma	0.1000	367	4	1
	0.1429	367	2	0
	0.2000	323	0	0
	0.3333	350	3	2
Gumbel	0.1028	298	4	0
	0.1828	325	0	0
	0.4112	325	3	2
	0.7311	349	20	9
		3949	64 (1.6%)	25 (0.6%)

test on errors of least squares fit for the three distributions chosen in this study.

The results of K-S test indicate that on the whole, the null hypothesis can be rejected with $\alpha=0.10$ sixty-four out of 3949 cases, or in about 1.6% while with $\alpha=0.05$ only 25 cases or about 0.6% must be rejected. Of all the samples at different levels of σ_k^2, Gumbel distribution with $\sigma_k^2=0.73$ showed relatively larger percent of rejections; 2.6% with $\alpha=0.05$ and 5.7% with $\alpha=0.10$. Discarding of negative variates generated with Gumbel distribution, which are about 8% at $\sigma_k^2=0.7$, might have led to relatively more irregular error terms and hence relatively more rejections of normality hypothesis.

The above results indicate that compared to the chi-square test the Kolmogorov-Smirnov test is considerably more favourable to the null hypothesis (that the errors are normal with zero mean). Since K-S test is more exact for samples of any size while the chi-square test is a large sample test it may be emphasized that the K-S test is more properly applicable to the small samples encountered in this work and preference to the conclusions drawn from it must be given. At the same time note may be taken of the fact that only in 14% of the cases the null hypothesis has been rejected by the chi-square test and thus the chi-square test also, in general, fails to reject the null hypothesis. The proper conclusion to be drawn from the above work thus seems to be that the hypothesis that the residual errors of the least squares fit (more specifically the fit by the least squares technique of the present work) are normally distributed with a zero mean and unknown variance σ^2 (the sample variance of error terms

gives an estimate of σ^2, Graybill (1961) which has been applied in this study) can not be rejected based upon the results obtained in this study. Thus the least squares method studied herein may be regarded as an application of the concepts of maximum likelihood.

Section V: Conclusions

The results of the study on least squares method presented in this appendix appear to substantiate the following observations and conclusions:

a) The use of least squares method in hydrology is theoretically sound, and may be regarded as an application of the concepts of Maximum likelihood.

b) The least squares estimates of parameters and percentiles are, in general, unbiased.

c) The least squares estimates of parameters and percentiles are, in general, less efficient than the maximum likelihood estimates.

d) Judicious choices of weighting functions may improve the efficiency of least squares estimates.

e) For the two parameter lognormal, gamma and Gumbel density functions, the particular choice of weights represented by the equation given below has improved the efficiency of least squares estimates and eliminated the bias of the estimates where observed.

$$w(x;\alpha,\beta) = f(x;\alpha,\beta)^{-\phi}$$

where $f(x;\alpha,\beta)$ represents the probability density function and has a value in the range of 0.75 to 1.00.

f) When applied to growing samples the least squares and maximum likelihood methods did not show any particular advantages over each other.

Section VI: A Suggested Procedure for use of Nonlinear Least Squares Method in Hydrology.

Section III presents the mathematics of nonlinear least squares method to estimate parameters of a two-parameter probabilistic model. By applying this method to fit LN, GA and GU, PDF's to their respective data it was found that

(i) When unitary weights were applied to the error terms i.e., $w_i = 1$ in Euqation A.30 ($\phi = 0.$ in Equations A.33 through A.41) the least squares method produced biased estimates of parameters in some cases, but no in all cases. The effect of this bias, in general, is to produce underestimates ($\sigma_F^2/S_k^2 < 1.0$) of hydrologic events (for different return periods).

(ii) Use of weight with $\phi = 0.25$ to 1.00 (Equations A.33 through A.41) virtually eliminated the bias mentioned in (i). However, use of weight in case of samples whose results were unbiased with unitary weights produced a bias in the parameter esitmates which resulted, in general, in over-prediction ($\sigma_F^2/S_k^2 > 1.0$) of hydrologic events (see Tables A.7 through A.9).

The above two results indicate that use of LS method in a specific form (i.e., with a recommended value of ϕ) can not be generalized. To obtain an unbiased solution by (weighted) LS method some value of ϕ in the range of $\phi = 0.0$ to 1.0 needs to be used in the (weighted) LS Equations. The specific value of ϕ which gives an unbiased LS solution can be determined only by a comparison of results with the parent PDF.

The two (extreme) cases of weighted LS, i.e., with $\phi = 0.0$ and with $\phi = 1.0$, are denoted in this thesis as LS method (since unitary weights are applicable to error terms) and MCS method (since with $\phi = 1.0$ the weighted

LS method reduces to the minimum chi-square' method described in Chapter II), respectively. Nevertheless, in view of results (i) and (ii) described above both LS and MCS methods may produce unbiased results in some cases and biased results in some cases. However, in general, the bias in LS produces under-estimates and the bias in MCS produces over-esitmates of hydrologic events, when data are fitted to the parent PDF. Thus, when applying least squares method the LS (ϕ = 0.0) and MCS (ϕ =1.00) should not to viewed as two separate methods, but should be viewed as the two extreme cases of the same method namely, weighted least squares, (WLS). The unbiased solution by WLS is obtained by trial and error (to determine the optimum ϕ value) in which LS and MCS may aid. In a real situation since populations are not known a priori, the following procedure is suggested to apply WLS to data samples, bith as a parameter estimation method, and as a method to discriminate PDF's.

(i) Assume that the criterion given by Equation 5.1 determines the parent PDF.

(ii) Choosing an appropriate PDF, compute the variance ratios $\sigma^2_{F,LS}/S^2_k$ and $\sigma^2_{F,MCS}/S^2_k$ (see Chapter IV for this notation) for the given data sample.

(iii) If the ratios computed in step (ii) differ greatly from 1.0 in the same direction the PDF chosen is not the 'best' PDF for the sample. The hydrologist/engineer may select another PDF and start again from step (ii).

(iv) If $\sigma^2_{F,LS}/S^2_k$ is less than 1.0 and $\sigma^2_{F,MCS}/S^2_k$ is greater than 1.0 the chosen PDF is probably the 'best' applicable to the sample. For this case the appropriate value of ϕ lies between 0.0 and 1.0 The solution by WLS may be determined by choosing ϕ such that

$\sigma^2_{F,WLS}/S^2_k$ is 1.0.

(v) If the two ratios computed in step (ii) differ from 1.0 in the same direction, but one of the ratios is close to 1.o the PDF chosen is possibly the 'best' PDF for the sample, but may not be the best. If both the ratios are less than 1.0, but $\sigma^2_{F,\ MCS}/S^2_k$ is closer to unity such a result might have occurred due to the positive bias of MCS method. On the other hand, if both the ratios are greater than 1.0, but $\sigma^2_{F,LS}/S^2_k$ is closer to unity such a reuslt might have occurred due to the negative bias of LS. At this stage results given by other PDF's as well as ML solution (where available) should be examined in making the selection of a PDF.

(vi) If the best PDF is not found in steps (iii) through (v) and the hydrologist/engineer wishes to choose one of the PDF's tried by him in step (iii) he may choose that PDF for which $\sigma^2_{F,LS/MCS}/S^2_k$ is closer to unity. (Note that, in this case either LS or MCS will give the solution for whish the variance ratio will be closer to unity).

APPENDIX B

THE TWO-PARAMETER LOGNORMAL DISTRIBUTION

The two-parameter lognormal distribution is defined by the following relation:

$$p(x;\ \mu_y,\ \sigma_y) = \frac{1}{x\ \sigma_y\sqrt{2\pi}}\, e^{-1/2\left[\frac{\ln x - \mu_y}{\sigma_y}\right]^2}\ ,\ x>0, \qquad \text{B.1}$$

where $y = \ln x$, μ_y = mean of y, and σ_y = standard deviation of y.

The lognormal distribution is positively skewed. The mean, variance, skewness coefficient and the kurtosis coefficient of this distribution are given in Table 3.1. For use in LS method(see Appendix A) the derivatives of p with respect to μ_y and σ_y may be written as

$$\frac{\partial p}{\partial \mu_y} = \left(\frac{\ln x - \mu_y}{\sigma_y^2}\right) p(x;\ \mu_y,\ \sigma_y) \qquad \text{B.2}$$

and

$$\frac{\partial p}{\partial \sigma_y} = \left[\frac{(\ln x - \mu_y)^2}{\sigma_y^3} - \frac{1}{\sigma_y}\right] p(x;\ \mu_y,\ \sigma_y) \qquad \text{B.3}$$

In the finite form, the lognormal distribution and its derivatives w.r.t. parameters μ_y and σ_y may be written, respectively as

$$\bar{p}(v_i;\mu_y,\ \sigma_y) = \frac{1}{\sigma_y\sqrt{2\pi}} \int_{v_{i-1}}^{v_i} \frac{1}{x}\, e^{-1/2\left[\frac{\ln x - \mu_y}{\sigma_y}\right]^2} dx \qquad \text{B.4}$$

$$\frac{\partial \bar{p}}{\partial \mu_y}(v_i;\mu_y,\ \sigma_y) = \int_{v_{i-1}}^{v_i} \left(\frac{\ln x - \mu_y}{\sigma_y^2}\right)\ p(x;\mu_y,\ \sigma_y)\ dx \qquad \text{B.5}$$

and

$$\frac{\partial \bar{p}}{\partial \sigma_y}(v_i;\mu_y,\ \sigma_y) = \int_{v_{i-1}}^{v_i} \left[\frac{(\ln x - \mu_y)^2}{\sigma_y^3} - \frac{1}{\sigma_y}\right]\ p(x;\mu_y,\ \sigma_y)\ dx \qquad \text{B.6}$$

Equations B.4 through B.6 are then easily evaluated using a numerical integration technique. In this work, the technique used was a simple trapezoidal rule with 16 sub-intervals on $(v_{i-1},\ v_i)$.

The moment estimates, $\hat{\mu}_{ym}$ and $\hat{\sigma}_{ym}$ of the parameters μ_y and σ_y are given by

$$\hat{\mu}_{ym} = 2\ln\bar{x} - 1/2\ \ln(SSM) \qquad \text{B.7}$$

and

$$\hat{\sigma}_{ym} = \sqrt{\ln(SSM) - 2\ \ln\bar{x}} \qquad \text{B.8}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $SSM = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

The maximum likelihood estimates $\hat{\mu}_{yml}$ and $\hat{\sigma}_{yml}$ of the parameters

μ_y and σ_y are given by

$$\hat{\mu}_{yml} = \frac{1}{n} \sum_{i=1}^{n} \ln x_i \qquad \text{B.9}$$

and

$$\hat{\sigma}_{yml} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\ln x_i)^2 - (\hat{\mu}_{yml})^2} \qquad \text{B.10}$$

For substitution in the LS normal equations(see Appendix A) the parameters μ_y and σ_y are treated as α and β, respectively.

APPENDIX C

THE TWO-PARAMETER GAMMA DISTRIBUTION

The two-parameter gamma distribution is defined by the relation

$$p(x;C,\ D) = \frac{C^D x^{D-1} e^{-Cx}}{\Gamma(D)},\ x > 0 \qquad C.1$$

$$= 0,\ \text{otherwise.}$$

where $\Gamma(D)$ is the gamma function defined by

$$\Gamma(D) = \int_0^\infty x^{D-1} e^{-x} dx,\ \text{for } D > 0 \qquad C.2$$

Integration of Equation C.1 leads to the recursive relationship

$$\Gamma(D) = (D - 1)\Gamma(D - 1) \qquad C.3$$

If β is a positive integer, then

$$\Gamma(D) = (D - 1)! \qquad C.4$$

In Equation C.1 the parameter D is called the shape parameter and C is called the scale parameter. There is a close relationship between the exponential distribution and gamma distribution. In fact, if D = 1.0, the gamma distribution reduces to the exponential distribution. The equations of mean, variance, skewness coefficient and the kurtosis coefficient of the gamma distribution are given by Table 3.1.

For use in LS method (See Appendix A.) the derivatives of p with respect to C and D may be written as

$$\frac{\partial p}{\partial C} = \left(\frac{D}{C} - x\right)\ p(x;\ C,D) \qquad C.5$$

and

$$\frac{\partial p}{\partial D} = \left(\frac{-\Gamma'(D)}{\Gamma(D)} + \ln C + \ln x\right)\ p(x;\ C,D). \qquad C.6$$

In the finite form, the gamma distribution and its derivatives

w.r.t. C and D may be written, respectively as

$$\bar{p}(v_i;\ C,\ D) = \frac{C^D}{\Gamma(D)} \int_{v_{i-1}}^{v_i} x^{D-1} e^{-Cx} dx, \qquad C.7$$

$$\frac{\partial \bar{p}}{\partial C}(v_i;\ C,\ D) = \int_{v_{i-1}}^{v_i} \left(\frac{D}{C} - x\right) p(x;\ C,\ D)\ dx \qquad C.8$$

and

$$\frac{\partial \bar{p}}{\partial D}(v_i;\ C,\ D) = \int_{v_{i-1}}^{v_i} \left\{\frac{-\Gamma'(D)}{\Gamma(D)} + \ln C + \ln x\right\} p(x;\ C,\ D)\ dx \qquad C.9$$

Equations C.7 through C.9 are then easily evaluated using a numerical integration technique. In this work, the technique used was a simple trapezoidal rule with 16 subintervals on $[v_{i-1},\ v_i]$.

Most computers have sub-routines to evaluate incomplete gamma functions.

To compute the derivatives of the gamma function, one first notes that

$$\frac{d}{dx}\left[\ln\Gamma(x)\right] = \frac{\Gamma'(x)}{\Gamma(x)} \qquad C.10$$

The function $\frac{\Gamma'(x)}{\Gamma(x)}$ is known as the Psi function, and is discussed in Abramowitz and Stegun (1964). It was given by the series expansion

$$\Psi(1 + z) = \frac{\Gamma'(1 + z)}{\Gamma(1 + z)} = -Eu + \sum_{n=1}^{\infty} \frac{z}{n(n+z)};\ z \neq -1, -2, \ldots \qquad C.11$$

Where Eu is Euler's constant (0.5772...). However, the infinite series in Equation C.11 converges very slowly and to obtain accuracy of six places, summation of on the order of a million terms would be required, an obvious impracticality. Based on a method by Kantorvich and Krylov (1958), an equivalent series having better convergence properties is

given by (Grant, 1973)

$$\sum_{n=1}^{\infty} \frac{z}{n(n+1)} = z(1.644394 - 1.202051z + 1.082323z^2 - z^3 \sum_{n=1}^{\infty} \frac{1}{n^4 (n+z)}) \qquad C.12$$

A similar handling of the terms of the expression (Abramovitz and Stegun, 1964) for Ψ',

$$\Psi'(z) = \sum_{j=1}^{\infty} \frac{1}{(z+j)^2} \quad (z \neq 0, -1, -2, \ldots) \qquad C.12$$

yields the equivalent series expression

$$\Psi'(z) = \frac{1}{z^2} + 1.644394 - 2.404102z + 3.2469699z^2 - z^3 \sum_{j=1}^{\infty} \frac{4j + 3z}{j^4 (j+z)^2} \qquad C.13$$

Equations C.12 and C.13 were used to compute the values of $\bar{\Psi}$ and Ψ' in this study. The first 100 terms of infinite series involved in Equations C.12 and C.13 were considered and the rest omitted.

The moment estimates $\hat{C}_m$ and $\hat{D}_m$ of the parameters C and D are given by

$$\hat{C}_m = \sum_{i=1}^{n} x_i \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad C.14$$

and

$$\hat{D}_m = \hat{C}_m \bar{x} \qquad C.15$$

Where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, (n = sample size).

The maximum likelihood estimates are computed as follows:

The likelihood function of a sample of size n would be

$$L = \frac{C^{nD}}{(\Gamma(D))^n} \prod_{i=1}^{n} x_i^{D-1} e^{-Cx_i} \qquad C.16$$

or $\ln L = nD \ln C - n \ln \Gamma(D) + (D - 1) \sum_{i=1}^{n} \ln x_i - C \sum_{i=1}^{n} x_i$.

Setting the derivatives of lnL w.r.t C and D equal to zero, one obtains

$$\frac{\partial \ln L}{\partial C} = \frac{nD}{C} - \sum_{i=1}^{n} x_i = 0$$

$$\text{or } C = D/\bar{x} \qquad \text{C.17}$$

and

$$\frac{\partial \ln L}{\partial D} = n \ln C - n \frac{\partial}{D} (\ln \Gamma(D)) + \sum_{i=1}^{n} \ln x_i = 0$$

$$\text{or } \ln C - \frac{\partial}{\partial D} (\ln \Gamma(D)) + \sum_{i=1}^{n} \frac{\ln x_i}{n} = 0$$

$$\text{or } \ln D - \ln \bar{x} - \frac{\partial}{\partial D} \ln \Gamma(D) + \sum_{i=1}^{n} \frac{\ln x_i}{n} = 0 \qquad \text{C.18}$$

Solving Equations C.17 and C.18 for C and D one obtains the ML estimates $\hat{C}_{ML}$ and $\hat{D}_{ML}$ of parameters C and D, respectively. However, no explicit solution for D is possible from Equation C.18. Starting from an initial value D_o (say, the moment estimate of D), Equation C.18 may be recursively solved by Newton's method to obtain an optimum value of D. The value thus obtained is treated as $\hat{D}_{ML}$ and $\hat{C}_{ML}$ is computed by

$$\hat{C}_{ML} = \hat{D}_{ML} \bar{x} \qquad \text{C.19}$$

For substitution in the LS normal equations (See Appendix A.) the parameters (C, D) of GA PDF are treated as (α, β).

APPENDIX D

THE TWO-PARAMETER GUMBEL DISTRIBUTION

The two parameter Gumbel or the type I asymptotic value distribution of largest vales, is defined by the relation

$$p(x;\ a,\ u) = ae^{-a(x-u)-e^{-a(x-u)}} \qquad -\infty \leq x \leq \infty \qquad \text{D.1}$$

In Equation D.1 the parameter 'a' is a measure of dispersion and is often called the shape or dispersion parameter. u is the mode of the distribution and also called the scale parameter.

The equations of the mean, variance, skewness coefficient and the kurtosis coefficient of Gumbel distribution are given by Table 3.1. For use in LS method (See Appendix A.) derivatives of p with respect to a and u may be written as

$$\frac{\partial p}{\partial a} = [(x - u)(e^{-a(x-u)} - 1) + \frac{1}{a}]p(x;\ a,\ u) \qquad \text{D.2}$$

and

$$\frac{\partial p}{\partial u} = a[1 - e^{-a(x-u)}]p(x;\ a,\ u). \qquad \text{D.3}$$

In the finite form, the Gumbel distribution and its derivatives w.r.t parameters a and u may be written, respectively, as

$$\overline{p}(v_i;\ a,\ u) = a\int_{v_{i-1}}^{v_i} e^{-a(x-u)-e^{-a(x-u)}}\,dx, \qquad \text{D.4}$$

$$\frac{\partial \overline{p}}{\partial a}(v_i;\ a,\ u) = \int_{v_{i-1}}^{v_i} [(x-u)(e^{-a(x-u)} - 1) + \frac{1}{a}]p(x;\ a,\ u)\,dx, \qquad \text{D.5}$$

and

$$\frac{\partial \overline{p}}{\partial u}(v_i;\ a,\ u) = \alpha \int_{v_{i-1}}^{v_i} [1 - e^{-a(x-u)}] p(x;\ a,\ u)\ dx \qquad D.6$$

Equations D.4 through D.6 are then easily evaluated using a numerical integration technique. In this work, the technique used was a simple trapezoidal rule with 16 sub-intervals on $(v_{i-1},\ v_i)$.

The moment estimates $\hat{a}_m$ and $\hat{u}_m$ of the parameters a and u are given by

$$\hat{a}_m = \frac{1.282}{\sqrt{SSM - (\overline{x})^2}} \qquad D.7$$

and

$$\hat{u}_m = x - \frac{0.5772}{\hat{a}} \qquad D.8$$

where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$SSM = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

No attempt has been made to evaluate the maximum likelihood estimates of parameters a and u as the iterative procedures involved would add considerably to the limited computer time available for this study. Harter and Moore (1967) obtained an explicit expression for the maximum likelihood estimator of location parameter u and the exact distribution of the estimator by assuming that the scale parameter a is known.

For substitution in the LS normal equations (See Appendix A.) the parameters (a, u) of GU PDF are treated as (α, β).

APPENDIX E

GENERATION OF SYNTHETIC VARIATES

In this appendix are discussed certain techniques by which samples of independent random variables with the given frequency distribution are generated. The basic approach to simulate samples of independent random variates having a certain frequency distribution consists of transforming the independent numbers of a uniform distribution on interval (0,1) to reproduce the required frequency distribution. The method is graphically explained by Figure E.1.

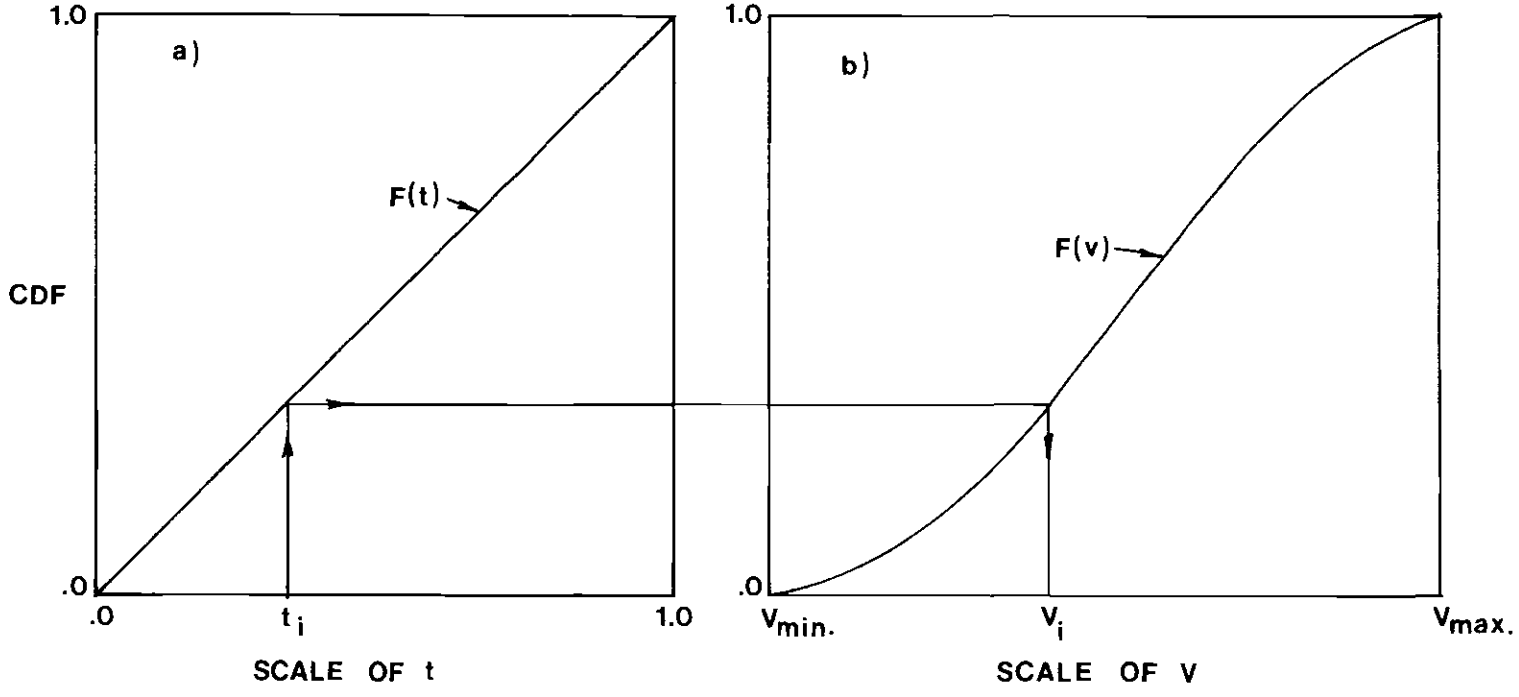


Figure E.1 Transformation of Uniform Random Numbers into Random Numbers of PDF, f(v)

The cumulative distribution functions, (CDF) of the uniform distribution and the frequency distribution to which the uniformly distributed random variables are to be transformed are represented by Figure E.1a) and Figure E.1b), respectively. The principle is that the probabilities of v_i and t_i are equal, or that they have one to one correspondence (Yevjevich, Stochastic Processes in Hydrology, 1972). It may be noted that all possible outcomes of t_i have equal probability of occurrence. The transformation is achieved by the relation

$$P(t \leq t_i) = P(v \leq v_i) \qquad (E.1)$$

that is, by equating CDF of t_i to that of v_i. Since the CDF of t_i is numerically equal to t_i itself (See Fig. E.1a) the relation given by Equation E.1 may be written as

$$t_i = F(v_i) \qquad (E.2)$$

The above transformation is usually expressed as an inverse transformation function given by

$$v_i = F^{-1}(t_i) \qquad (E.3)$$

Equation E.3 should be understood in the Sense of Equation E.2.

Equation E.3 is easily programmable on a digital computer for some frequency distributions.

From the foregoing paragraphs it is seen that to simulate random variates of a given frequency distribution a sequence of independent random numbers uniformly distributed on interval (0,1) is to be first generated. One obvious source of obtaining such random numbers is by observation of any random process in nature, such as emission of particles from radioactive material. However, such sources are not really satisfactory since the numbers generated would have to be recorded, then read into the computer and stored in valuable memory locations; the entire

process would be expensive. Instead, computer programs which use the arithmetic capabilities of a computer to generate uniformly distributed random numbers are not available. But, computers are deterministic machines in that, given identical inputs they produce identical results. A deterministic machine using a deterministic algorithm or computational scheme is impossible to generate truly "random" numbers. Computers can, however, generate "pseudorandom" numbers, which are sequences of numbers carefully (deterministically) constructed to maintain the important properties of truly random sequences (Fiering and Jackson, 1971).

One procedure for generating pseudorandom uniformly distributed numbers by digital computers is the mixed linear congruential method that operates as follows. Select some starting value $X_o \geq 0$, a multiplier $a \geq 0$, an increment $C \geq 0$ and a modulus (or divisor) m which is larger than X_o, a, and C. Use X_o as the first element in the sequence. Let

$$X_1 = (aX_o + C) \bmod (m) \qquad (E.4)$$

where the 'mod m' notation means to take the quantity aX_o+C, subtract m as many times as possible without driving the result negative and then set X, equal to the result. In an equivalent form X is the remainder. When aX_o+C is divided by m. At each subsequent step X_n is generated from the previous value X_{n-1} by the formula

$$X_n = (aX_{n-1}+C) \bmod (m) \qquad (E.5)$$

If C=0 in Equation E.5, the generator is called linear congruential.

Knuth (1969) discusses rules that should be applied in selecting a, m and C values so as to make the sequence as random as possible. In case of binary computers m is set equal to 2^ρ, where ρ is the word length of the machine in question. Thus, the pseudorandom numbers generated will be uniformly distributed on interval $(0,2^\rho-1)$. Usually 2^ρ is an extra-

ordinarily large number (typically ρ=32, 2^{32} = 4,294,967,296) and the probability of repetition of a sequence of pseudorandom numbers generated is very remote. It is this fact that makes pseudorandom number generators usable.

The specific procedures used to generate independent random synthetic variates of the three frequency distributions studied in this work (i.e., lognormal, gamma, and Gumbel) are described below.

Lognormal Random Synthetic Variates

It may be recalled that if the random variable X is lognormally distributed, then the transformed random variate Y=lnX, is normally distributed with mean μ_y and standard deviation σ_y. Conversely, if Y is normally distributed with mean μ_y and standard deviation σ_y, the transformed variate X=Exp(Y) is lognormally distributed. This property was used in this work to generate lognormal synthetic variables.

Let $y \sim N(\mu_y, \sigma_y^2)$

and t be uniformly distributed on (0,1)

Now, ϕ, the standard normal cumulative distribution function is given by

$$\phi(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} \exp(-Z^2/2)\, dz \qquad \text{(E.6)}$$

By the relation given by E.3, one obtains

$$y = \sigma_y\, \phi^{-1}(t) + \mu_y \qquad \text{(E.7)}$$

and the lognormally distributed variate, X is simply given by

$$X = \exp(y) \qquad \text{(E.8)}$$

All numerical steps through Equation E.7 in the above procedure are computed by the subprogram RANDN of UNIVAC 1108 for given μ_y and σ_y. This subprogram first obtains uniformly distributed pseudorandom numbers on $(0, 2^{35}-1)$ from the NRAND routine and divides them by 2^{35} to convert them

uniformly random on (0,1). Then it uses a STAT-PACK function subprogram TINORM to accomplish inverse normal distribution (See UNIVAC Large Scales Systems Stat-Pack, 1970).

The inverse of the normal cumulative distribution function (Equation E.6) can not be expressed in closed form. The following approximation was made in the subprogram TINORM (Abramovity and Stegun, 1964).

$$y = p - \frac{a_o+a_1p+a_2p^2}{1+b_1p+b_2p^2+b_3p^3} \qquad \text{(E.9)}$$

where

$p = \ln 1/t^2$ if $0<t\leq 0.5$

$p \quad \ln 1/(1-t)^2$ if $.5<t<1$

$a_o = 2.515517$

$a_1 = 0.802853$

$a_2 = 0.010328$

$b_1 = 1.432788$

$b_2 = .189269$

$b_3 = .001308$

GAMMA RANDOM SYNTHETIC VARIATES

The inverse of gamma cumulative distribution function can not be expressed in closed form for use in equation E.3.

An approximate formula by which random numbers of gamma population with parameters α and β (see Appendix C), with β integral, may be generated is given by (Naylor, T.H., et al., 1966)

$$v_i = -\frac{1}{\alpha}\left(\ln \prod_{k=1}^{\beta} t_{ik}\right) \qquad \text{(E.10)}$$

where t_{ik} form a double sequence of numbers uniform and random on (0,1).

Equation E.14 was used in this study to generate pseudorandom gamma variates. To generate uniformly distributed pseudorandom numbers on (0,1) for use in Equation E.10 a subprogram RANDUJ was written. The subprogram RANDUJ obtains, first, pseudorandom numbers uniformly distributed on $(0,2^{35}_{-1})$ by MRAND subroutine and divides them by 2^{35} to make them uniformly distributed pseudorandom numbers on (0,1).

GUMBEL RANDOM SYNTHETIC VARIATES

The Cumulative Distribution Function CDF, of Gumbel distribution (see Appendix D) is given by

$$F(v) = \exp\left[-e^{-\alpha(v-w)}\right] \qquad \text{(E.11)}$$

Let γ be the reduced variate, given by $\gamma = \alpha(v-u)$ (E.12)

The CDF of Gumbel distribution in terms of reduced variate, γ, is given by

$$F(\gamma) = \exp(-e^{-\gamma}) \qquad \text{(E.13)}$$

In Equation E.13, the value of γ can be easily evaluated for the known value of $F(\gamma)$. The known values of $F(\gamma)$ are t_i, i=1,2...n which are uniformly distributed random numbers on (0,1). Having evaluated

γ_i for given t_i the Gumbel random variates with parameters α and u may be computed by the formula

$$V_i = \frac{\gamma i}{\alpha} + u \qquad \text{(E.14)}$$

In this work uniform pseudorandom numbers on (0,1) obtained from subprogram RANDU of UNIVAC 1108 were transformed to Gumbel pseudorandom variates of parameters α and u by Equations E.13 and E.14. The subprogram RANDU first obtains uniform pseudorandom numbers on $(0,2^{35}-1)$ by RANDU routine and divides them by 2^{35} to convert them uniformly random on (0,1).

Since Gumbel distribution has $-\infty$ as its lower limit it is likely that some negative values of V_i may be generated. The programming was done such that the negative V_i's generated were always discarded and the generation procedure continued until the sample contained the required number of positive variates. The number of negative values generated with each sample was recorded.

'GROWING' a SAMPLE

To 'grow' a sample, by which it was meant that the larger sample retained the data in the smaller sample, the pseudorandom numbers were generated as sets of variates included in the largest sample. The first n variates of each set are then used as the sample. For example, assume that samples of initial size 25 are to be 'grown' to samples of sizes 50,75 and 100, respectively. The procedure consists of generating sets of 100 variates and choosing the first 25, 50, 75 of the 100 members of each set as the samples. Thus, it was insured that data in larger samples contained the data of smaller samples.

APPENDIX F

A DESCRIPTION OF THE COMPUTER PROGRAM

In this appendix is given a user oriented description of the computer program developed for the least squares (LS) maximum likelihood (ML) and moments (MO) fitting of Lognormal (LN), gamma (GA), and Gumbel (GU) (ML is not available for this PDF) PDF's (all two parameter) to a given data sample.

Purpose of the Program

This program is designed to accept real samples or generate data samples of LN, GA or GU distributions and fit the data samples to a LN, GA or GU distribution by the methods of MO, ML (not available for GU) and LS. The program also evaluates predicted values for the return periods 10, 25, 50, 100, 200, 500 and 1000 years by ML and LS methods for LN and GA fits and by LS and MO methods for GU fit. The program performs chi-square and Kolmogorov-Smirnov goodness-of-fit tests for LS fit and also performs certain statistical tests on the Least Squares errors terms and on the parameter estimates.

Language and Computer Requirements

This program is coded in Fortran V and was designed for operation on the UNIVAC 1108 under EXEC 8 monitor. The program requires 12719 decimal locations in the instruction bank and 15484 decimal locations in the data bank. I/O is by the standard input and output devices only. No temporary or permanent files or storage are required for operation of the program. Use is made in the program of elements of the UNIVAC Large scale systems MATH-PACK/STAT-PACK program group.

Data Input Format

Input in all cases is according to the following format (see Chapter V for the description of a run):

Card	Format	Variables
1,2,3	16A5	TITLE1
4	2I5	IOPT,JOPT,
5	5I5	ITER,NCYCLE,NRETR, NEMPT,NPE
6	7F10·2	RETP
7	7F6·4	PE
8	7F10·6	RETV
9	5F10·3	PAR(1), PAR(2), TEST, WEIGHT, GSF
10	3F10·4	RNI,WSCALE,PVA

The following cards are required if IOPT=0

11-1	8X,I3,17A4	NX,TITLE
11-2	8X,11F6·0	VT(NX)

Repeat 11-2 until all VT(NX) are accommodated

Repeat 11-1 and 11-2 NCYCLE number of times for the NCYCLE number of Real data samples.

The above variables have the following meanings:

TITLE1 - Description of the run (3 cards)

IOPT - a control variable

IOPT=0 Reads Real data

1 Generates Lognormal variates

2 Generates Gamma variates

3 Generates Gumbel variates

JOPT - a control variable

JOPT - 1 Lognormal PDF is fit to data

2 Gamma PDF is fit to data

3 Gumbel PDF is fit to data

NX - sample size (For synthetic data NX=100xGSF

ITER - the maximum number of iterations per cycle for the iteration of least squares procedure. If convergence of the least squares procedure is not obtained within ITER iterations, the case is abandoned and computations begun on the following case.

NCYCLE - the number of synthetic data samples to be generated or the number of real samples input.

NRETP - the number of return periods for which predictions are to be evaluated ($\leq$ 7)

NEMPT - the number of empty classes to be added to the data histogram of LS fit (NEMPT=0 for the present study)

NPE - NPE=NRETP

RETP - Return periods in years NRETP number of values are to be input

PE - Probability of exceedence corresponding to each return period under RETP NRETP number of values, one for each RETP in the same order are to be input.

RETV - The population predictions for each return period under RETP. NRETP number of values, one for each RETP in the same order are to be input. (Leave RETV blank if IOPT=0)

PAR(1), PAR(2)- α and β, respectively of the population selected (LN, GA or GU). Leave PAR(1) and PAR(2) blank if IOPT=0.

TEST - the limiting value of the parameter corrections in least squares fit. If $(\Delta\alpha^2 + \beta^2)^{1/2} \leq$ TEST, convergence is declared.

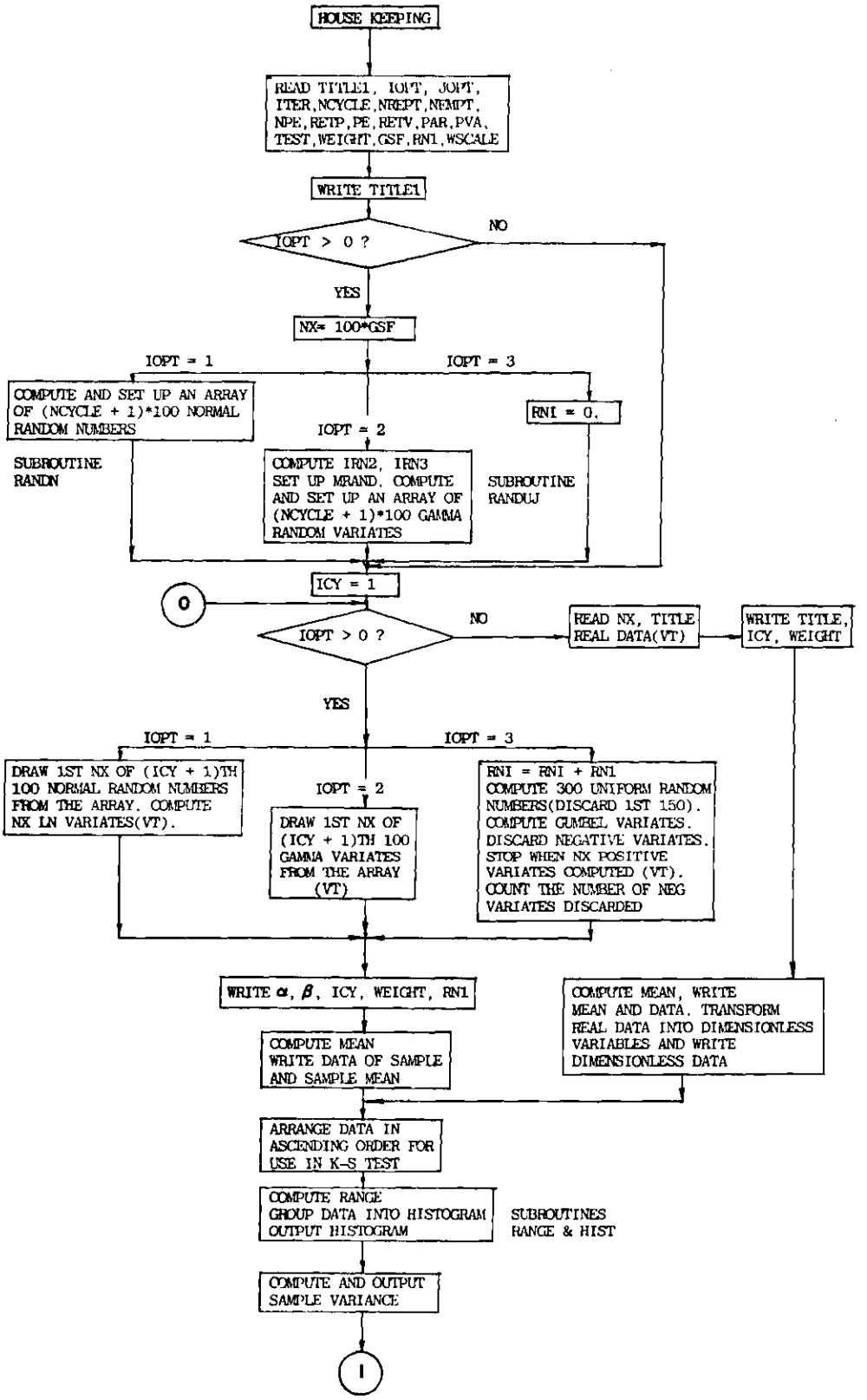


Figure F.1 Flow Chart of Main Program

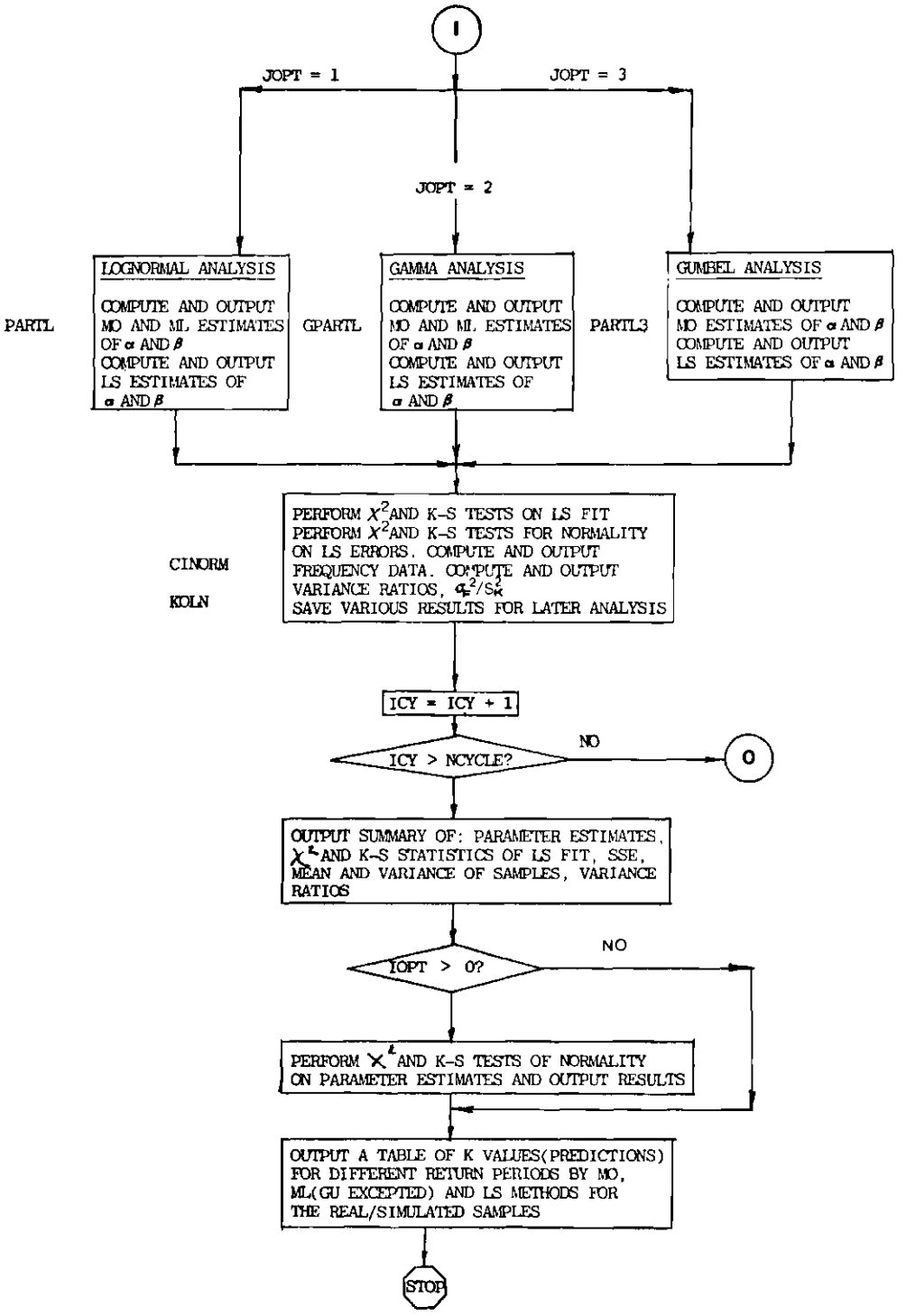
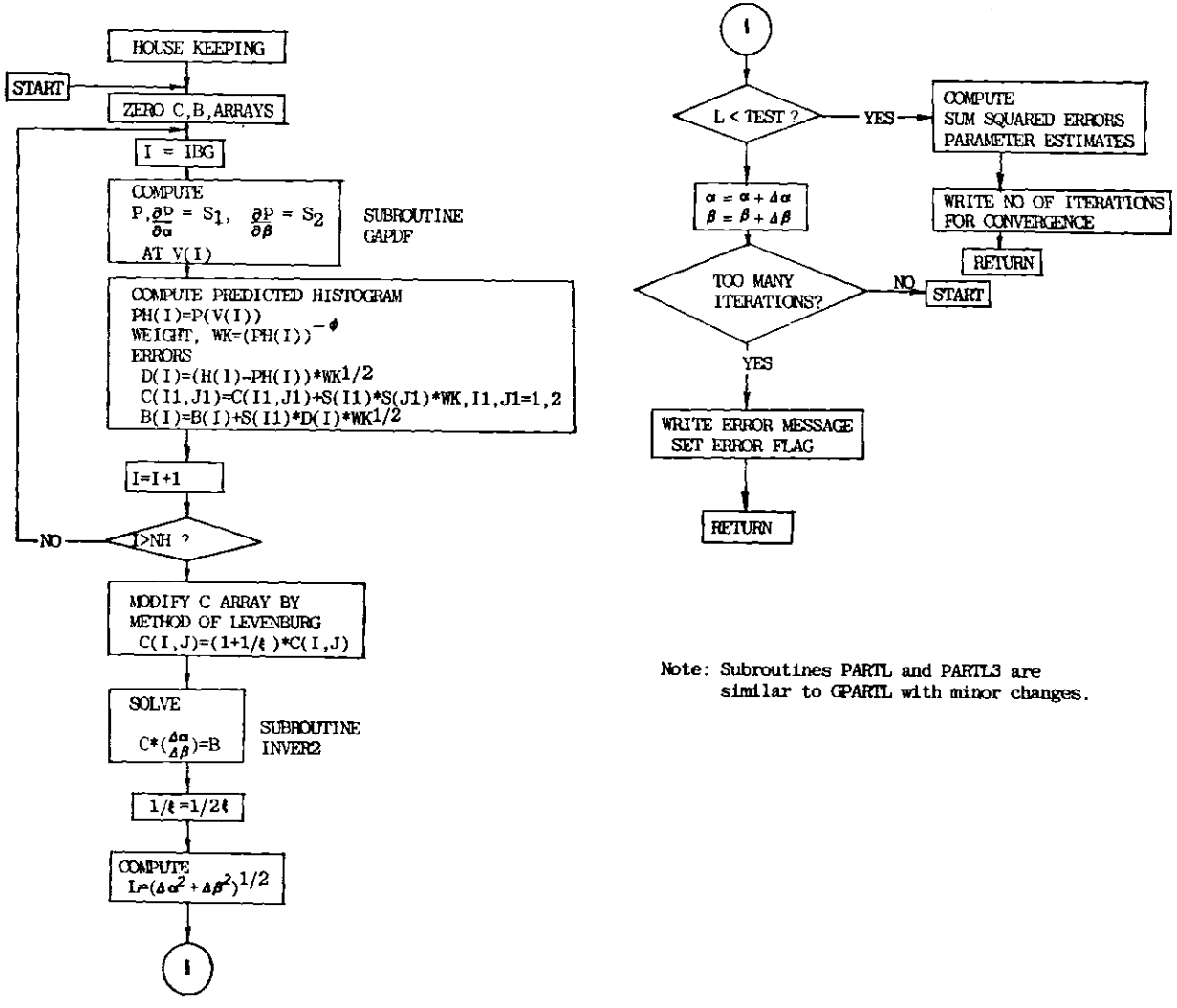


Figure F.1 -Continued



Note: Subroutines PARTL and PARTL3 are similar to GPARTL with minor changes.

Figure F.2 Flow Chart of Subroutine GPARTL

WEIGHT - the exponent ϕ of the weight factor in the expression $[P(v_i; \alpha,\beta)]^{-\phi}$.

GSF - Growing Sample Factor; the factor by which 100 is to be multiplied to obtain samples containing the first (100)x (GSF) of data items of samples with 100 as sample size (Note: $0 < GSF \leq 1.0$. If the sample size is 100, GSF = 1.0. Leave GSF blank if IOPT=0.)

RN1 - the initial number from which the synthetic data are generated (Leave blank if IOPT=0).

WSCALE - the correction factor by which Sturges' class interval is to be multiplied (see Appendix A, Section -)

PVA - the population variance of synthetic data (LN, GA or GU) (Leave blank if IOPT=0)

TITLE - the title for real data sample

VT - values of real data to be analyzed. NX number of values are to be input.

Flow charts of the main program, which directs the logical flow of the program, and of subroutine GPARTL, which performs the least squares fitting of GA density function to the histogram, are given in Figure F.1 and F.2, respectively. A list of other subroutines required (exclusive of those subroutines in the standard FORTRAN library) and a brief description of their methods and functions is given in Table F.1.

Output and run time vary with the size of samples being analyzed. Run time should average between two to three seconds of computer CPU time per case for normal samples of size 100 or less.

TABLE F.1. DESCRIPTION OF COMPUTER SUBROUTINES

Subroutine	Calling Program	Function and Methods
NRAND*	MAIN	The auxilary random number generator (see Appendix E).
RANDN*	MAIN	Generates normally distributed random numbers
RANDUJ	MAIN	Generates uniform random numbers on (0,1) by the use of the methods of Appendix E.
RANGE*	MAIN	Computes the range of a sample.
HIST*	MAIN	Groups a given set of data into a histogram and prints the histogram on the printer.
PARTL	MAIN	Performs least squares fitting of LN density function to the histogram (see Appendix A, Section).
PARTL3	MAIN	Performs least squares fitting of GU density function to the histogram (see Appendix A, Section).
PSI	MAIN	Computes $\frac{d}{dx}(\ln\Gamma(x))$ (see Appendix E).
PSIP	MAIN	Computes $\frac{d^2}{dx^2}(\ln\Gamma(x))$ (see Appendix E).

TABLE F.1. DESCRIPTION OF COMPUTER SUBROUTINES

Subroutine	Calling Program	Function and Methods
GAMIN*	MAIN	Evaluates the incomplete gamma function.
CINORM	MAIN	Performs a chi-square test for normality (see Appendix A, Section).
KOLN	MAIN	Performs a Kolmogorov-Smirnov test for normality (see Appendix A, Section).
LNPDF	PARTL	Evaluates the lognormal distribution and its derivatives (see Appendix B).
GAPDF	GPARTL	Evaluates the gamma distribution function and its derivatives (see Appendix C).
GUPDF	PARTL3	Evaluates the Gumbel distribution and its derivatives (see Appendix D).
INVER2	PARTL, GPARTL, PARTL3	Inverts a 2X2 matrix.
GAMMA*	GAPDF	Evaluates the complete gamma function (see Appendix B).
GROUP*	HIST	Groups data into a histogram.

TABLE F.1. DESCRIPTION OF COMPUTER SUBROUTINES

Subroutine	Calling Program	Function and Methods
PLOT1*	HIST	Plots a line of symbols on the printer.
RNORM*	KOLN	Evaluates the cumulative distribution function of the normal distribution.
MRAND*	RANDUJ	Generates integers random on $(0, 2^{35}-1)$
TINORM*	CINORM	Evaluates the inverse of the cumulative distribution function of a normal distribution.
CHI*	CINORM	Evaluates the cumulative distribution function of the chi-square distribution.

* denotes subroutines included in the Univac large scale systems MATH-PACK/ STAT-PACK group.

APPENDIX G

DESCRIPTION OF COMPUTER RUNS

In this appendix is given a list of the various simulation runs made during the course of this study. Tables G-1, G-2 and G-3 summarize the runs for samples of LN, GA and GU PDF's, respectively. Data samples of each run series have the same population parameters (i.e., the same population variance, σ_k^2). Table 5.1 gives the values of population parameters and population variance for each run series. The other parameters of the runs consisted of the sample size n, the value of weight exponent ϕ of the weighted LS method (Appendix A), the fitted PDF and the initial number RN1 from which the pseudo-random numbers were generated by the computer. All runs consisted of 25 samples of size n.

In the following tables the runs consisting of the same data samples were generally designated by a single run number and the specific combinations of ϕ-n-fitted PDF' used in individual runs are given under 'Run Variation'. ϕ=0.00 indicates a LS fit and ϕ=1.00 indicates the MCS method. When n is less than 100 the samples consist of the 'first' n data items of the samples with size 100 of the run with the same run number.

Table G.1. Parameters of the Simulation Runs - LN Data

RUN SERIES	RUN NUMBER	RN1	RUN VARIATION	FITTED PDF	ϕ	n
1LN	1LN1	12345.	A	LN	0.00	100
			B	LN	0.25	100
			C	LN	0.50	100
			D	LN	0.75	100
	1LN2	312345.	A	LN	0.00	100
			B	LN	0.25	75
			C	LN	0.50	50
			D	LN	0.75	25
	1LN3	57.	A	GA	0.00	100
	1LN4	712345.	A	GA	0.00	100
			B	GU	0.25	100
			C	GU	0.50	100
			D	LN	0.75	100
	1LN5	712345.	A	LN	0.00	100
			B	LN	1.00	100
			C	LN	0.00	100
			D	LN	0.00	100
			E	LN	0.00	100
			F	LN	0.00	100
			G	LN	1.00	100
			H	LN	0.00	100
			I	LN	1.00	100
2LN	2LN1	23456.	A	LN	0.00	100
			B	LN	0.25	100
			C	LN	0.50	100
			D	LN	0.75	100
	2LN2	323456.	A	LN	0.00	100
			B	LN	0.25	100
			C	LN	0.50	100
			D	LN	0.75	100
			E	LN	0.00	75
			F	LN	0.00	50
			G	LN	0.00	25
	2LN3	723456.	A	LN	0.00	100
			B	LN	0.25	100
			C	LN	0.50	100
			D	LN	0.75	100
	2LN4	333333.	A	LN	0.00	100
	2LN5	11111.	A	LN	0.00	100
			B	LN	1.00	100
			C	GA	0.00	100
			D	GA	1.00	100
			E	GU	0.00	100
			F	GU	1.00	100

Table G.1. Parameters of the Simulation Runs - LN Data(Continued)

RUN SERIES	RUN NUMBER	RN1	RUN VARIATION	FITTED PDF	ϕ	n
3LN	3LN1	34567.	A	LN	0.00	100
			B	LN	0.25	100
			C	LN	0.50	100
			D	LN	0.00	75
			E	LN	0.00	50
			F	LN	0.00	25
	3LN2	334567.	A	LN	0.00	100
			B	LN	0.25	100
			C	LN	0.50	100
			D	LN	0.75	100
	3LN3	734567.	A	LN	0.00	100
			B	LN	0.25	100
			C	LN	0.50	100
			D	LN	0.75	100
	3LN4	111111.	A	LN	0.00	100
			B	LN	1.00	100
			C	GA	0.00	100
			D	GA	1.00	100
			E	GU	0.00	100
			F	GU	1.00	100
4LN	4LN1	345678.	A	LN	0.00	100
			B	LN	0.25	100
			C	LN	0.50	100
			D	LN	0.75	100
			E	LN	1.00	100
	4LN2	45678.	A	LN	0.00	100
			B	LN	0.25	100
			C	LN	0.50	100
			D	LN	0.75	100
	4LN3	111111.	A	LN	0.00	100
	4LN4	13579.	A	LN	0.00	100
			B	LN	1.00	100
			C	LN	0.00	75
			D	LN	0.00	50
			E	LN	0.00	25
			F	GA	0.00	100
			G	GA	1.00	100
			H	GU	0.00	100
			I	GU	1.00	100

Table G.2. Parameters of the Simulation Runs - GA Data

RUN SERIES	RUN NUMBER	RN1	RUN VARIATION	FITTED PDF	ϕ	n
1GA	1GA1	1234.	A	GA	0.00	100
			B	GA	0.50	100
			C	GA	0.75	100
	1GA2	11234.	A	GA	0.00	100
			B	GA	0.50	100
			C	GA	0.75	100
	1GA3	111555.	A	GA	0.00	100
			B	GA	1.00	100
	1GA4	21234.	A	GA	0.00	100
	1GA5	2345.	A	GA	0.00	100
			B	GA	1.00	100
			C	GA	0.00	75
			D	GA	0.00	50
			E	GA	0.00	25
			F	LN	0.00	100
			G	LN	1.00	100
			H	GU	0.00	100
			I	GU	1.00	100
2GA	2GA1	12345.	A	GA	0.00	100
			B	GA	0.75	100
	2GA2	333777.	A	GA	0.00	100
			B	GA	1.00	100
	2GA3	11111.	A	GA	0.00	100
	2GA4	22345.	A	GA	0.00	100
	2GA5	2345.	A	GA	0.00	100
			B	GA	1.00	100
			C	GA	0.00	75
			D	GA	0.00	50
			E	GA	0.00	25
			F	LN	0.00	100
			G	LN	1.00	100
			H	GU	0.00	100
			I	GU	1.00	100

Table G.2. Parameters of the Simulation Runs - GA Data(Continued)

RUN SERIES	RUN NUMBER	RN1	RUN VARIATION	FITTED PDF	ϕ	n
3GA	3GA1	3456.	A	GA	0.00	100
			P	GA	0.75	100
	3GA2	13456.	A	GA	0.00	100
			B	GA	0.75	100
	3GA3	23456.	A	GA	0.00	100
			B	GA	0.75	100
	3GA4	555999.	A	GA	0.00	100
			B	GA	1.00	100
	3GA5	23456.	A	GA	0.00	100
	3GA6	2345.	A	GA	0.00	100
			B	GA	1.00	100
			C	GA	0.00	75
			D	GA	0.00	50
			E	GA	0.00	25
			F	LN	0.00	100
			G	LN	1.00	100
			H	GU	0.00	100
			I	GU	1.00	100
4GA	4GA1	14567.	A	GA	0.00	100
			B	GA	0.75	100
	4GA2	214567.	A	GA	0.00	100
			B	GA	0.75	100
	4GA3	4567.	A	GA	0.00	100
			B	GA	0.75	100
	4GA4	779911.	A	GA	0.00	100
			B	GA	1.00	100
	4GA5	2345.	A	GA	0.00	100
			B	GA	1.00	100
			C	GA	0.00	75
			D	GA	0.00	50
			E	GA	0.00	25
			F	LN	0.00	100
			G	LN	1.00	100
			H	GU	0.00	100
			I	GU	1.00	100

Table G.3. Parameters of the Simulation Runs - GU Data

RUN SERIES	RUN NUMBER	RN1	RUN VARIATION	FITTED PDF	ϕ	n
1GU	1GU1	333333.	A	GU	0.00	100
			B	GU	1.00	100
	1GU2	77.	A	GU	0.00	100
	1GU3	555777.	A	GU	0.00	100
	1GU4	739789.	A	GU	0.00	100
	1GU5	777777.	A	GU	0.00	100
			B	GU	1.00	100
			C	GU	0.00	75
			D	GU	0.00	50
			E	GU	0.00	25
			F	LN	0.00	100
			G	LN	1.00	100
			H	GA	0.00	100
			I	GA	1.00	100
2GU	2GU1	555555.	A	GU	0.00	100
			B	GU	1.00	100
	2GU2	77.	A	GU	0.00	100
	2GU3	555777.	A	GU	0.00	100
	2GU4	987987.	A	GU	0.00	100
			B	GU	1.00	100
			C	LN	0.00	100
			D	LN	1.00	100
			E	GA	0.00	100
			F	GA	1.00	100
	2GU5	777777.	A	GU	0.00	100
			B	GU	1.00	100
			C	GU	0.00	75
			D	GU	0.00	50
			E	GU	0.00	25
	2GU6	333333.	A	GU	0.00	100
			B	GU	1.00	100

Table G.3. Parameters of the Simulation Runs - GU Data(Continued)

RUN SERIES	RUN NUMBER	RN1	RUN VARIATION	FITTED PDF	ϕ	n
3GU	3GU1	333333.	A	GU	0.00	100
			B	GU	1.00	100
	3GU2	33.	A	GU	0.00	100
	3GU3	555777.	A	GU	0.00	100
	3GU4	897897.	A	GU	0.00	100
	3GU5	777777.	A	GU	0.00	100
			B	GU	1.00	100
			C	GU	0.00	75
			D	GU	0.00	50
			E	GU	0.00	25
			F	LN	0.00	100
			G	LN	1.00	100
			H	GA	0.00	100
			I	GA	1.00	100
	3GU6	555555.	A	GU	0.00	100
			B	GU	1.00	100
4GU	4GU1	999999.	A	GU	0.00	100
			B	GU	1.00	100
	4GU2	777777.	A	GU	0.00	100
	4GU3	555777.	A	GU	0.00	100
	4GU4	789.	A	GU	0.00	100
	4GU5	77.	A	GU	0.00	100
			B	GU	1.00	100
			C	GU	0.00	75
			D	GU	0.00	50
			E	GU	0.00	25
			F	LN	0.00	100
			G	LN	1.00	100
			H	GA	0.00	100
			I	GA	1.00	100

APPENDIX H

SUMMARY OF REAL DATA

In this appendix is given a summary of real data used in this study for various purposes. The real data used in this study consist of annual peak flows from 67 stream gauging stations located throughout the United States. These data were taken from U.S. Geological Survey Water Supply Papers Numbers 1671 through 1689 and Water Resources Data for various states published by Geological Survey.

Table H.1 shows the 67 stream gauging stations in the variance order of data. Columns a through d in Table H.1 represent the following:

a — Serial number in the variance order of data samples.

b,c — The station part number and station number (inventory numbers) used by Geological Survey, respectively.

d — Years of flow record. The first two digits represent the beginning year and the last two digits represent the ending year of flow record used in the analysis.

TABLE H.1:

LIST OF STREAM GAGING STATIONS

a	b	c	d	Gauging Stations
1	6A	375	1 14 70	MADISON RIVER NEAR W. YELLOWSTONE, MONT
2	9	2395	1 04 70	YAMPA RIVER AT STEAMBOAT SPRINGS, COLO
3	4	770	1 12 69	WOLF RIVER AT KESHENA, WISC
4	3A	510	1 08 70	TYGARD VALLEY RIVER AT BELINGTON, W VA
5	4	2525	1 11 67	BLACK RIVER NEAR BOONVILLE, NY
6	3A	115	1 05 71	ALLEGHENY RIVER AT RED HOUSE, NY
7	10	1285	1 05 70	WEBER RIVER NEAR OAKLEY, UTAH
8	9	850	1 06 70	ROARING FORK AT GLENWOOD SPRINGS, COLO
9	10	1685	1 01 63	BIG COTTONWOOD CREEK NEAR SALT LAKE CITY, UTAH
10	1A	450	1 03 71	DEAD RIVER AT THE FORKS, MAINE
11	3A	1835	1 96 72	GREENBRIAR RIVER AT ALDERSON, W VA
12	1B	5405	1 00 72	SUSQUEHANNA RIVER AT DANVILLE, PA
13	4	735	1 98 72	FOX RIVER AT BERLIN, WISC
14	1B	3210	1 12 67	SACANDAGA RIVER NEAR HOPE, NY
15	3A	155	1 10 70	BROKENSTRAW CREEK AT YOUNGSVILLE, PA
16	9	470	1 11 70	BLUE RIVER AT DILLON, COLO
17	1B	5480	1 11 70	N. BALD EAGLE CREEK AT BEECH CR. STA., PA
18	3A	205	1 10 70	OIL CREEK AT ROUSEVILLE, PA
19	2A	195	1 96 73	JAMES RIVER AT BUCHANAN, VA
20	3A	325	1 10 70	REDBANK CREEK AT ST. CHARLES, PA
21	3A	215	1 11 70	FRENCH CREEK AT CARTERS CORNERS, PA
22	4	2165	1 13 67	LITTLE TONAWANDA CREEK AT LINDEN, NY
23	1A	315	1 03 73	PISGATAQUIS RIVER NEAR DOVER-FOXCROFT, MAINE
24	14	3210	1 06 73	UMPQUA RIVER NEAR ELKTON, OREG
25	10	1700	1 99 63	MILL CREEK NEAR SALT LAKE CITY, UTAH
26	6B	7070	1 09 70	N FORK S PLATTE RIVER AT S PLATTE, COLO
27	14	2100	1 09 70	CLACKAMAS RIVER NEAR CAZADERO, OREG
28	1E	4340	1 04 67	DELAWARE RIVER AT PORT JERVIS, NY
29	2A	835	1 06 71	TAR RIVER AT TARBORO, NC
30	2B	4790	1 05 73	PASCAGOULA RIVER AT MERRILL, MISS
31	2B	2235	1 94 73	OCONEE RIVER AT DUBLIN, GA

TABLE H.1:

LIST OF STREAM GAUGING STATIONS (continued)

a	b	c	d	Gauging Stations
32	2B	3350	1 03 70	CHATTAHOOCHEE RIVER NEAR NORCROSS, GA
33	2B	3920	1 92 73	ETOWAH RIVER AT CANTON, GA
34	4	1560	1 10 70	TITTABAWASSEE RIVER AT MIDLAND, MICH
35	6A	625	1 15 70	TENMILE CREEK NEAR RIMINI, MONT
36	13	3190	1 04 70	GRANDE RIVER RONDE RIVER AT LA GRANDE, OREG
37	7	725	1 05 68	BLACK RIVER AT BLACK ROCK, ARK
38	2E	4820	1 09 71	PEARL RIVER AT EDINBURG, MISS
39	2B	3495	1 05 70	FLINT RIVER AT MONTEZUMA, GA
40	2A	550	1 97 73	ROANOKE RIVER AT ROANOKE, VA
41	2B	4415	1 93 73	TOMBIGBEE RIVER AT COLUMBUS, MISS
42	4	1130	1 01 72	GRAND RIVER AT LANSING, MICH
43	8	660	1 03 69	TRINITY RIVER AT RIVERSIDE, TX
44	1B	6385	1 95 66	POTOMAC RIVER AT POINT OF ROCKS, MD
45	1B	3615	1 11 67	CATSKILL CREEK AT OAK HILL, NY
46	1A	940	1 10 70	SOUHEGAN RIVER AT MERRIMACK, N.H.
47	5	4645	1 03 65	CEDAR RIVER AT CEDAR RAPIDS, IOWA
48	2B	4770	1 05 73	CHICKASAWHAY RIVER AT ENTERPRISE, MISS
49	5	3310	1 67 72	MISSISSIPPI RIVER AT ST. PAUL, MINN
50	5	4815	1 05 65	DES MOINES RIVER NEAR BOONE, IOWA
51	2B	2185	1 04 70	OCONEE RIVER NEAR GREENSBORO, GA
52	1B	3345	1 11 67	HOOSIC RIVER NEAR EAGLE BRIDGE, NY
53	11	5025	1 17 70	WILLIAMSON RIVER NEAR CHILOQUIN, OREG
54	5	145	1 13 70	SWIFTCURRENT CREEK AT MANY GLACIER, MONT
55	11	1520	1 06 72	ARROYO SECO NEAR SOLEDAD, CALIF
56	1B	6680	1 08 73	RAPPAHANNOCK RIVER NEAR FREDRICKSBURG, VA
57	8	335	1 04 69	NECHES RIVER NEAR ROCKLAND, TEX
58	14	3590	1 06 73	ROGUE RIVER AT RAYGOLD NEAR CENT. POINT, OREG
59	10	1720	1 02 63	EMIGRATION CREEK NEAR SALT LAKE CITY, UTAH
60	9	4060	1 10 71	VIRGIN RIVER AT VIRGIN, UTAH
61	11	2665	1 17 70	MERCED RV. AT POHONO BR. NEAR YOSEMITE, CALIF
62	11	2820	1 17 70	MID. TUOLUMNE RV. AT OAKLAND REC. CAMP, CAL
63	11	2750	1 16 70	FALLS CREEK NEAR HETCH HETCHY, CALIF

TABLE H.1:

LIST OF STREAM GAUGING STATIONS (continued)

a	b	c	d	Gauging Stations
64	1A	1805	1 11 72	MIDDLE BR. WESTFIELD RV. AT GROSS HEIGHTS, MASS
65	11	4095	1 11 69	OREGON CREEK NEAR NORTH SAN JUAN, CALIF
66	11	2035	1 02 65	TULE RIVER NEAR PORTERVILLE, CALIF
67	11	980	1 14 72	ARROYO SECO NEAR PASADENA, CALIF

BIBLIOGRAPHY

1. Abramowitz, Milton and Stegun, Irene A (editors) : Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables, National Bureau of Standards, Applied Mathematics Series 55, Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., June 1976.

2. Bowker, Albert H. and Lieberman, Gerald J. : Engineering Statistics, Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1972.

3. Booth, G. W. and Peterson, T. I. : Non-linear Estimation, I.B.M. Share Program. Pa., No. 687 WLNL1 (1958).

4. Bulletin No. 13 : Methods of Flow Frequency Analysis, Sub committee on Hydrology, Inter-Agency committee on Water Resources, Washington, D.C., April 1966.

5. Bulletin No. 15 : A Uniform Technique for Determining Flood Flow Frequencies, Water Resources Council, Washington, D.C., 1967.

6. Bard, Yonathan : Nonlinear Parameter Estimation, Academic Press, 1974.

7. Benjamin, Jack R. and Cornell, C. Allin : Probability, Statistics, and Decision for Civil Engineers, McGraw-Hill Book Co., New York, 1970.

8. Beard, L. R., Flood Flow Frequency Techniques, Center for Research in Water Resources, The University of Texas at Austin, 1974.

9. Chow, Ven Te. : Handbook of Applied Hydrology. McGraw-Hill, New York, 1964.

10. Cramer, Harold : Mathematical Methods of Statistics, Princeton University Press, Princeton, New Jersey, 1946.

11. Davies, O. L. (Editor) : Design and Analysis of Industrial Experiments, Oliver and Boyd, Ltd., Edinburgh, Scotland, 1954.

12. Decoursey, Donn G. and Snyder Willard M. : Computer Oriented Method of Optimizing Hydrologic Model Parameters, Journal of Hydrology, Vol. 9, 1969, pp. 34-36.

13. Draper, N. R. and Smith, H. : Applied Regression Analysis, John Wiley & Sons, Inc., New York, 1966.

14. Fiering, M. B. and Jackson, Barbara B. : Synthetic Streamflows, Water Resources Monograph 1, American Geophysical Union, Washington, D.C., 1971.

15. Fisher, R. A. : On an Absolute Criterion for Fitting Frequency Curves, Messenger of Math, Vol. 41, p. 155, 1912.

16. Grant, James L. : Statistical Frequency Analysis by Optimization of Density Functions to Histograms, Ph.D. Thesis, Georgia Institute of Technology, Atlanta, November 1973.

17. Graybill, Franklin A. : An Introduction to Linear Statistical Models, McGraw-Hill Book Co., New York, 1961.

18. Halperin, Max : Confidence Interval Estimation in Non-linear Regression, Sperry Rand Research Center, Sudberry, Mass., 1962.

19. Hartley, H. O. : "The Modified Gauss-Newton Method for the Fitting of Nonlinear Regression Functions by Least Squares," Technometrics, Vol. 2, No. 2, 1961, pp. 269-280.

20. Hartley, H. O. : "Exact Confidence Regions for the Parameters in Non-linear Regression Laws," Biometrica, Vol. 51, University College, London, G.B. 1964, pp. 347-353.

21. Hartley, H. O. and Booker, Aaron : Nonlinear Least Squares Estimation, Annals of Mathematical Statistics, Vol. 36, 1965, pp. 638-650.

22. Harter, H. L., and A. H. Moore : A Note on Estimation from a Type I Extreme-Value Distribution, Technometrics, Vol. 9, No. 2, pp. 325-331, May 1967.

23. Hines, William W. and Montgomery, Douglas C : Probability and Statistics In Engineering and Management Science, The Ronald Press Co., New York, 1972.

24. Kantorovich, L. V. and Krylou, V. I. : Approximate Methods of Higher Analysis, translated by Curtis D. Benster, Inter-science Publishers, Inc., New York, 1958.

25. Kendall, M. G. and Stuart, A. : The Advanced Theory of Statistics, Vol. 1 and II, Charles Griffin and Co., Ltd., 42 Drury Lane, London, England, 1973.

26. Knuth, Donald E. : The Art of Computer Programming, Vol. 2, Seminumerical Algorithms, Addison-Wesley, 1969.

27. Levenberg, Kenneth : "A Method for the Solution of Certain Non-linear Problems in Least Squares," Quarterly of Applied Mathematics, Vol. 2, No. 2, 1944, pp. 164-168.

28. Lindgren B. W. and McElrath, G. W. : Introduction to Probability and Statistics, Second Edition, The Macmillan Company, New York, 1966.

29. Kennedy, John B. and Neville, Adam M. : Basic Statistical Methods for Engineers and Scientists, A Dun-Donnelley Publisher, New York, 1976.

30. Markovic, R. D. : Probability Distributions of Best Fit to Distributions of Annual Precipitation and Runoff, Hydrology Paper No. 8, Colorado State University, Fort Collins, Colorado, 1965.

31. Marquardt, D. W. : An Algorithm for Least Squares Estimation of Nonlinear Parameters, Journal of the Society of Industrial and Applied Mathematics, Vol. 11, No. 2, June 1963, pp. 431-441.

32. Nahi, Nasser E. : Estimation Theory and Applications, John Wiley & Sons, Inc. 1969.

33. Natrella, Mary Gibbons : Experimental Statistics, Handbook 91, National Bureau of Standards, U.S. Department of Commerce, U.S. Government Printing Office, Washington, D.C., 1963.

34. Naylor, T. H., et al., Computer Simulation Techniques, John Wiley & Sons, Inc., New York, 1966.

35. Rao, C. R. : Linear Statistical Interference and Its Application, John Wiley & Sons, Inc., New York, 1965.

36. Rao, D. V. : A Study on Monthly Streamflow Simulation, Master's Special Research Problem (unpublished), School of Civil Engineering, Georgia Institute of Technology, Atlanta, 1974.

37. Robey, Donald L. and Wallace, J. R. : A Study of Selected Flood Frequency Methods, unpublished report, School of Civil Engineering, Georgia Institute of Technology, Atlanta, Georgia, August, 1969.

38. Schulz, E. F., et al., (editors) : Floods and Droughts, Proceedings of the Second International Symposium in Hydrology, Water Resources Publications, Fort Collins, Colorado, 1973.

39. Snyder, Willard M.: "Fitting of Distribution Functions by Non-linear Least Squares," Water Resources Research, Vol. 8, No. 6, December, 1972.

40. Snyder, Willard M. : "Some Possibilities for Multivariate Analysis in Hydrologic Studies," Journal of Geographical Research, Vol. 67, No. 2, February, 1962.

41. Snyder, Willard M. and Wallace, James R. : Fitting a Three-Parameter Log-Normal Distribution by Least Squares, Nordic Hydrology 5, 1974.

42. Sturges, Herbert A. : The Choice of a Class Interval, Journal of the American Statistical Association, Vol. 21, 1926, page 65.

43. Taylor, Angus E. : Advanced Calculus, Quinn and Company, Boston, Mass., 1955.

44. Univac Large Scale Systems Math-Pack, Sperry Rand Corporation, UP 7542, Rev. 1, 1972.

45. Univac Large Scale Systems Stat-Pack, Sperry Rand Corporation, UP 7502, Rev. 1, 1970.

46. Von Mises, Richard : Mathematical Theory of Probability and Statistics, Academic Press, New York, 1964.

47. Wasan, M. T. : Parametric Estimation, McGraw-Hill Book Co., New York, 1970.

48. Yevjevich V. : Probability and Statistics in Hydrology, Water Resources Publications, Fort Collins, Colorado, U.S.A., 1972.

49. Yevjevich, V. : Stochastic Processes in Hydrology, Water Resources Publications, Fort Collins, Colorado, U.S.A., 1972.

VITA

Donthamsetti Veerabhadra Rao was born on June 5, 1937, in Ramachandrapuram, East Godavari District, the State of Andhra Pradesh, India. He is the son of late Donthamsetti Chandriah and Donthamsetti Subbamma. (It is customary for the citizens of the State of Andhra Pradesh, to place their family name in front of the given name i.e., the first name. The author has preserved this native tradition. Thus, 'Donthamsetti' is the author's family name while 'Veerabhadra Rao' forms his first name). He received Bachelor's and Master's degrees in Civil Engineering from the Indian Institute of Technology, Kharagpur in the years 1959 and 1960, respectively. He then worked as a teacher for twelve years in India (Senior Fellow, Technical Teachers' Training Program 1960-62. Lecturer in Dam Construction, Irrigation and Hydraulics, College of Engineering, Poona, 1962-66. Head of the Department of Applied Mechanics, Government Polytechnic, Khamgaon, 1966-72). He attended the Georgia Institute of Technology since 1972, received the degree of Master of Civil Engineering in 1974 and worked toward the degree of Doctor of Philosophy in Civil Engineering. He was married in 1964 to the former Miss Kalepu Sreedevi of Kakinada, Andhra Pradesh. He has three sons.